N° ordre : 09/2006-M/EL

REPUBLIQUE ALGERIENNE DEMOCRATIQUE ET POPULAIRE
MINISTERE DE L'ENSEIGNEMENT SUPERIEUR ET DE LA RECHERCHE SCIENTIFIQUE

UNIVERSITE DES SCIENCES ET DE LA TECHNOLOGIE HOUARI BOUMEDIENE

FACULTE D'ELECTRONIQUE ET D'INFORMATIQUE

THESE

Présentée pour l'obtention du diplôme de MAGISTER

En : ELECTRONIQUE

Spécialité : ELECTRONIQUE DES SYSTEMES

Par **Abderrezak HALILALI**

Thème

Analyse et transformation du signal de parole par rapport à un locuteur cible

Soutenue le : 11 / 11 / 2006, devant le jury composé de :

Mr Y. SMARA	Professeur	USTHB	Président
Mr A . HOUACINE	Professeur	USTHB	Directeur de thèse
Mme M. BOUDRAA	Professeur	USTHB	Examinatrice
Mr A. DJERADI	Maître de conférence	USTHB	Examinateur

Remerciements

Je voudrais pour commencer exprimer ma profonde gratitude à Monsieur A. HOUACINE, qui a su me redonner confiance et qui m'a suivi tout le temps nécessaire pour la réalisation de ce travail, avec la patience qui le caractérise.

Je tiens également à remercier les membres du jury qui ont accepté sans hésitations à évaluer mon travail : Le président du jury Pr Y. SMARA et les examinateurs :
Pr. M.BOUDRAA et Pr A. DJERADI.

Que Hammoud MERRAD trouve ici le témoignage de ma gratitude pour son soutient pendant les moments les plus critiques.

Je voudrais également remercier tous les collègues qui n'ont pas cessé de me « bousculer » affectueusement à finaliser ce travail, qu'ils se reconnaissent ici pour leur aide précieuse.

Ma gratitude va à mon épouse qui a su être patiente et à mes enfants que j'ai du quelque peu délaisser ces derniers temps.

Je dédie cette thèse à mon père et à ma défunte mère
ainsi qu'a ma mon épouse et à mes enfants raison de tous les efforts consentis.

TABLE DES MATIERES

Chapitre 4 : Outils de transformation

Chapitre 5 : Mise en oeuvre de la conversion de voix

TABLE DES FIGURES

LISTE DES ABREVIATIONS

AR	Auto-Régressif
CELP	Code-Excited Linear Prediction (excitation codée en prédiction linéaire)
DFW	Dynamic Frequency Warping (programmation dynamique fréquentielle)
DTW	Dynamic Time Warping (programmation dynamique temporelle)
HMM	Hidden Markov Model
LAR	Log-Area Ratio (logarithme des rapports d'aires)
LPC	Linear Predictive Coding (codage par prédiction linéaire)
LSF	Line Spectral Frequency
LSP	Line Spectral Pair (paires de lignes spectrales)
MBE	Multi Band Excited (excitation multi-bande)
MFCC	Mel-Frequency Cepstrum Coefficients (coefficients cepstraux sur l'échelle de Mel)
MOS	Mean Opinion Score
PSOLA	Pitch Synchronous OverLap Add (addition-recouvrement synchrone au pitch)
RELP	Residual Excited Linear Prediction (excitation par le résidu en prédiction linéaire)
RI	Réponse Impulsionnelle
SNR	Signal To Noise Ratio (rapport signal à bruit)
TF	Transformée de Fourier
TFCT	Transformée de Fourier Court Terme
TFD	Transformée de Fourier Discrète
VQ	Vector Quantization (quantification vectorielle)

INTRODUCTION

Ces dernières années ont été témoins d'avancées rapides dans la technologie de signal de parole avec un nombre croissant de produits qui emploient le signal de parole en tant que moyen d'interaction homme-machine à l'exemple du TTS (Text To Speech) et le GSM. Ces résultats ne sont pas le fruit du hasard, ils sont dus à la collaboration efficace établie entre les chercheurs.

L'identification, la reconnaissance et les TTS (Text To Speech) ont eu la priorité des efforts de recherches dirigés vers l'interaction homme-machine et des solutions efficaces ont émergé avec un souci permanent d'amélioration du naturel des signaux de parole.

Les domaines de recherche en traitement de la parole aussi variés que la modélisation, le codage, l'identification, la synthèse, la linguistique et autres, sont autant de domaines qui ont contribué à cette émergence et partagent beaucoup de méthodes et approches communes. Il y a également plusieurs axes de recherche qui servent de point de raccordement entre ces champs principaux, la conversion de voix est l'une d'entre elles. Elle combine les méthodes d'analyse du signal de parole et l'extraction des règles d'identification avec les méthodes de transformation et synthèse de signal de parole à la lumière de la perception et de la linguistique.

C'est donc dans ce cadre général de l'analyse-transformation-synthèse que s'inscrit notre travail.

Pour la transformation des signaux, les méthodes permettant de réaliser une modification de l'échelle temporelle ou fréquentielle consistent en les méthodes paramétriques, reposant sur un modèle de signal précis (par exemple, modèle sinusoïdal), et les méthodes non-paramétriques (où il n'est fait aucune hypothèse sur la nature du signal traité).
Celles-ci peuvent se répartir en deux catégories : méthodes travaillant dans le domaine temporel et méthodes opérant dans le domaine fréquentiel.

Les deux modèles que nous considérerons sont des modèles de signaux : le modèle PSOLA ou Pitch Synchronous Overlap-Add, modèle non-paramètrique, et le modèle classique LPC (modèle paramétrique). Ces deux modèles correspondent à deux interprétations différentes des signaux. Dans le cas d'un signal harmonique, le premier considère le signal comme la répétition périodique d'une forme d'onde, le second comme la sortie d'un filtre linéaire variant dans le temps qui simule les caractéristiques de la fonction de transfert du conduit vocal.

Les applications de ces méthodes sont multiples :
Les techniques de modification de l'échelle temporelle trouvent des applications dans le domaine de l'enseignement des langues (possibilité de modifier la vitesse d'élocution) ou du contrôle de bande audio (possibilité d'accélérer un enregistrement en maintenant l'intelligibilité de la parole) alors que les techniques de

modification de l'échelle fréquentielle sont utilisées principalement pour la synthèse de la parole à partir du texte pour le bon rendu de la prosodie.

Dans notre travail nous nous sommes intéressés dans un premier temps aux méthodes permettant de réaliser indépendamment une modification de l'échelle temporelle ou fréquentielle d'un signal. Puis à leur application à la conversion de voix.

La transformation de voix est une opération qui consiste à modifier les enregistrements audio d'une voix afin d'en changer l'identité perçue. On peut par exemple vouloir créer un enregistrement de voix d'homme à partir d'une voix de femme, en changer le timbre, la durée, l'intonation, l'accent ou la prononciation sur certaines régions, etc.

La conversion de voix s'inscrit dans ce contexte en se fixant plus particulièrement comme objectif d'imiter une voix précise (dite cible) en transformant les enregistrements d'une voix source.

S'il est important d'utiliser des algorithmes de synthèse de bonne qualité, afin que les signaux générés paraissent naturels et qu'aucun artéfact ne vienne les dégrader, le point central de la procédure est la fonction de transformation des paramètres d'analyse en fonction de connaissances sur les voix source et cible.
Ces connaissances proviennent d'une étape préalable d'apprentissage des différences des voix selon les descripteurs caractéristiques de leur identité.

On utilisera l'extraction, l'apprentissage et la modification de paramètres du signal vocal (fréquence fondamentale, formants, excitation, etc.) liés au locuteur.

A cette fin, nous commençons au chapitre 1 par établir quels sont ces descripteurs et quelles sont leurs correspondances avec les paramètres bas niveau manipulés lors d'une analyse-synthèse.
Le chapitre 2 abordera la notion de transformation du signal de parole par la méthode PSOLA (Pitch Synchronus OverLap and Add)et ses variantes.
Dans le chapitre 3 nous définirons la conversion de voix et nous y décrirons un état de l'art sur les fonctions de transformation.
Au chapitre 4, nous présenteront les outils principaux permettant de procéder à une conversion de voix ; nous y traiterons de l'alignement des signaux de voix, préalable à l'apprentissage lui-même ; ainsi que d'autres outils comme les méthodes de détection de pitch et marquage.
La synthèse de ces éléments se fait au chapitre 5 en présentant dans son ensemble la procédure de transformation retenue et les implémentations associées. Nous discuterons des résultats obtenus, des améliorations et des voies à suivre pour un travail futur.

Chapitre 1

Caractéristiques de l'identité vocale

1.Caractéristiques de l'identité vocale.

De nombreux chercheurs se sont penchés sur le problème de l'étude des caractéristiques acoustiques de la voix sans parvenir jusqu'à présent à déterminer parfaitement les paramètres qui définissent l'identité vocale d'un locuteur.
Les paragraphes qui suivent présentent succinctement les observations et les résultats de recherches dans ce domaine.

1.1 Production de la parole.

Le processus de production de la parole peut être représenté par le modèle source-filtre[Fant70].Le signal de parole est modélisé comme la sortie d'un filtre linéaire variant dans le temps, qui simule les caractéristiques spectrales de la fonction de transfert du conduit vocal, excité par un signal source qui reflète l'activité des cordes vocales dans les zones voisées et le bruit de friction dans les zones non voisées. Bien que simpliste, cette représentation est capable de décrire la majorité des phénomènes de la parole
et a été à la base de nombreux codeurs et synthétiseurs de parole.

La décomposition source/filtre est une théorie particulièrement bien adaptée au problème de la conversion de voix. Transformer les paramètres du filtre revient à simuler la modification des caractéristiques du conduit vocal alors que la modification des paramètres du signal source simule les changements de la prosodie et des caractéristiques du signal d'excitation glottique.

1.2 Discrimination de locuteurs

1.2.1 Les problèmes de variabilité de la parole.

La parole est un phénomène a priori très simple à comprendre. Tout un chacun n'est-il après tout pas capable de suivre une conversation ? Mais l'homme peut rencontrer des difficultés lorsqu'il essaie de suivre une conversation dans une langue autre que sa langue maternelle, même s'il la connaît bien. Et que dire, et que comprendre surtout, lorsqu'il essaie de suivre une conversation dans une langue qui lui est inconnue ! Ce dernier cas est

pourtant le plus proche du problème posé en reconnaissance automatique de la parole (RAP), la machine n'ayant aucune connaissance propre en compréhension de la parole.

- **Variabilité intra-locuteur**

La variabilité intra-locuteur identifie les différences dans le signal produit par une même personne. Cette variation peut résulter de l'état physique ou moral du locuteur. Une maladie des voies respiratoires peut ainsi dégrader la qualité du signal de parole de manière à ce que celui-ci devienne totalement incompréhensible, même pour un être humain. L'humeur ou l'émotion du locuteur peut également influencer son rythme d'élocution, son intonation ou sa phraséologie.
Il existe un autre type de variabilité intra-locuteur lié à la phase de production de parole ou de préparation à la production de parole. Cette variation est due aux phénomènes de coarticulation. Il est possible de voir la phase de production de la parole comme un compromis entre une minimisation de l'énergie consommée pour produire des sons et une maximisation des scores d'atteinte des cibles que sont les phonèmes tels qu'ils sont théoriquement définis par la phonétique.
Un locuteur adoptera donc un compromis qui est généralement partagé par une vaste majorité de la communauté de langage à laquelle il appartient.
La variabilité intra-locuteur est cependant beaucoup plus limitée que la variabilité inter-locuteur

- **Variabilité inter-locuteur**

La variabilité inter-locuteur est un phénomène majeur en reconnaissance de la parole. Comme nous venons de le rappeler, un locuteur reste identifiable par le timbre de sa voix malgré une variabilité qui peut parfois être importante. La contrepartie de cette possibilité d'identification à la voix d'un individu est l'obligation de donner aux différents sons de la parole une définition assez souple pour établir une classification phonétique commune à plusieurs personnes.
La cause principale des différences inter-locuteurs est de nature physiologique. La parole est principalement produite grâce aux cordes vocales qui génèrent un son à une fréquence

de base, le fondamental. Cette fréquence de base sera différente d'un individu à l'autre et plus généralement d'un genre à l'autre, une voix d'homme étant plus grave qu'une voix de femme, la fréquence du fondamental étant plus faible. Ce son est ensuite transformé par l'intermédiaire du conduit vocal, délimité à ses extrémités par le larynx et les lèvres. Cette transformation, par convolution, permet de générer des sons différents qui sont regroupés selon les classes que nous avons énoncées précédemment. Or le conduit vocal est de forme et de longueur variables selon les individus et, plus généralement, selon le genre et l'âge. Ainsi, le conduit vocal féminin adulte est, en moyenne, d'une longueur inférieure de 15% à celui d'un conduit vocal masculin adulte. Le conduit vocal d'un enfant en bas âge est bien sûr inférieur en longueur à celui d'un adulte. Les convolutions possibles seront donc différentes et, le fondamental n'étant pas constant, un même phonème pourra avoir des réalisations acoustiques très différentes.

La variabilité interlocuteur trouve également son origine dans les différences de prononciation qui existent au sein d'une même langue et qui constituent les accents régionaux.

La variabilité interlocuteur telle qu'elle vient d'être présentée permet de comprendre aisément pourquoi les méthodes de reconnaissance des formes fondées sur la quantification de concordances entre une forme à analyser et un ensemble de définitions strictes plus ou moins formelles ne peuvent être appliquées, avec un succès limité, qu'à des applications où le nombre de définitions est restreint.

Mais la variabilité interlocuteur, malgré son importance évidente, n'est pas encore la variabilité la plus importante car les différences au sein des classes phonétiques sont en nombre restreint. L'environnement du locuteur est porteur d'une variabilité beaucoup plus importante.

- **Variabilité due à l'environnement**

La variabilité liée à l'environnement peut, parfois, être considérée comme une variabilité intra-locuteur mais les distorsions provoquées dans le signal de parole sont communes à toute personne soumise à des conditions particulières. La variabilité due à l'environnement peut également provoquer une dégradation du signal de parole sans que le locuteur ait modifié son mode d'élocution.

La variabilité environnementale due au locuteur peut tout d'abord être de nature physiologique.

Ainsi, un système mécanique provocant une déformation du conduit vocal provoquera immanquablement une variation dans le signal de parole produit. Ces contraintes physiques sont généralement rencontrées dans les systèmes de transport où une posture particulière, ou une accélération lors du déplacement, pourront provoquer une déformation.

Enfin, le stress et l'angoisse que certaines personnes finissent par éprouver lors de longs voyages peuvent également être mis au rang des contraintes environnementales susceptibles de modifier le mode d'élocution.

1.2.2 Niveaux d'observation.

Il ressort de ce qui précède qu'il n'est pas évident d'identifier les paramètres acoustiques qui jouent un rôle décisif sur la caractérisation de l'identité vocale. L'identité vocale est un mélange entre divers paramètres acoustiques dont le degré et l'ordre d'importance différent d'un individu à l'autre.

Ces paramètres peuvent être classés en trois catégories selon les niveaux d'observation adoptés.

- **Niveau segmental**

Au niveau segmental, les caractéristiques du signal de parole sont analysées à l'échelle centi-seconde. Les descripteurs acoustiques des paramètres segmentaux incluent les formants et leurs largeurs de bande, la pente spectrale, la fréquence fondamentale et l'énergie. Les paramètres segmentaux dépendent principalement des propriétés physiologiques et physiques des organes de la parole, mais également de l'état émotionnel du locuteur.

- **Niveau supra-segmental**

Les paramètres supra-segmentaux décrivent la variation des paramètres segmentaux sur des échelles d'observation plus grande que le segment acoustique.

Ils incluent les contours de pitch, les trajectoires spectrales, les durées des phonèmes, et l'évolution de l'énergie sur une expression. Ces paramètres dépendent généralement du style d'élocution du locuteur et de son état psychologique.
Pour traiter les paramètres supra-segmentaux, des modèles complexes doivent être utilisés pour caractériser les paramètres prosodiques et acoustiques à l'échelle de l'énoncé pour chacun des locuteurs et d'autre part pour formaliser le lien entre ces locuteurs.

- **Niveau linguistique.**

Au niveau linguistique, d'autres types d'informations peuvent également être considérées comme porteurs de l'identité vocale : l'univers lexical du locuteur, la phonétisation propre au locuteur ou encore d'autres caractéristiques relevant de dialectes ou d'accents régionaux. Ces phénomènes linguistiques sont au-delà de la portée de cette thèse et ne seront pas étudiés dans le présent document.

Les chercheurs ont prouvé que les paramètres segmentaux et supra-segmentaux sont perceptuellement significatifs pour l'identification du locuteur. Parmi les paramètres supra-segmentaux, les contours de la fréquence fondamentale sont d'une importance particulière. Parmi les paramètres segmentaux, les chercheurs ont considéré que l'enveloppe spectrale et notamment les paramètres de formant sont d'importance majeure. Une approche de conversion de voix tenant compte d'un ensemble plus complet de paramètres acoustiques est susceptible de surpasser toutes les approches basées sur un ensemble plus réduit de paramètres acoustiques.

1.2.3 Discrimination des locuteurs par des humains.

Dans le domaine de l' identification du locuteur qui est un problème très proche de la conversion de voix, la recherche de paramètres caractérisant le locuteur ont amené Matsumoto et al. [MHSN73] à conclure que la *fréquence fondamentale* est le paramètre le plus caractéristique de l'identité vocale.
En effet Matsumoto et al. ont étudié la corrélation entre les scores de reconnaissance du locuteur et certains paramètres acoustiques tels que les fréquences des trois premiers

formants, la pente du spectre de la source glottique, le pitch moyen, et les fluctuations rapides de la période de pitch. Les auteurs ont conclu que la fréquence fondamentale moyenne seule conduit à 55% de taux de reconnaissance du locuteur. En ajoutant la pente moyenne du spectre de la source glottique et la fluctuation de la période de pitch ce taux de reconnaissance passe à 63.8%. Si d'autre part, les fréquences des trois premiers formants sont ajoutées à la fréquence fondamentale moyenne seule, ce taux atteint 69,3%. Ainsi, ils ont conclu que la fréquence fondamentale moyenne joue un rôle important dans la perception de l'identité vocale, et que la contribution relative des caractéristiques du conduit vocal à l'identité du locuteur est plus grande que celle des caractéristiques de la source glottique. Tous ces paramètres ensemble contribuent à 86 % de taux identification du locuteur par l'audition.

Dans une étude similaire, Itoh et Saito [IS88] ont abouti à une conclusion différente. A l'aide d'un test ABX (un test dans lequel des auditeurs jugent si le son prononce par un locuteur "X" est plus proche du locuteur "A" ou de "B") les auteurs ont conclu que *l'enveloppe spectrale* du signal de parole est le facteur le plus déterminant en terme d'identification du locuteur par l'audition.

La plupart des techniques courantes de reconnaissance de locuteur sont basées sur la caractérisation de la distribution statistique des enveloppes spectrales.
Il paraît qu'on peut exactement distinguer des locuteurs par la comparaison de leurs enveloppes spectrales respectives. [IS88]
Si en reconnaissance du locuteur, les chercheurs s'intéressent aux paramètres acoustiques permettant à un ordinateur de distinguer entre différents locuteurs, l'enjeu en conversion de voix est la perception par les humains.

Et si les techniques de reconnaissance automatique de la parole reposent essentiellement sur la comparaison de paramètres acoustiques évalués a travers des distribution statistiques, cela est du au fait que l'identification ou la reconnaissance du locuteur est tranchée selon un critère de seuil de reconnaissance ou d'identification à atteindre .
Par contre dans les systèmes de synthèse de la parole à partir du texte (TTS : Texte To Speech), de transformation ou de conversion de voix, la finalité étant une perception par les humains ; la fréquence fondamentale associée à l'enveloppe spectrale semble jouer un rôle prépondérant dans la discrimination des locuteurs.

Chapitre 2

Modification du signal de parole

2. Modification du signal de parole

2.1 Définitions.

La modification de l'échelle temporelle permet d'altérer arbitrairement la durée d'un signal sans en modifier le contenu fréquentiel. Bien sûr, la lecture du signal à une fréquence d'échantillonnage différente de celle de l'enregistrement conduit bien à une modification de la durée du signal, mais altère également son contenu fréquentiel (ex: disque 45 tours joué en 33 tours), ce que l'on cherche à éviter.

La modification de l'échelle fréquentielle ("pitch shifting") est l'opération duale de la précédente, et consiste à modifier la hauteur d'un son donné, sans en modifier la durée.

On peut distinguer deux types de transposition:

- **Le changement de hauteur sans modification des formants.**

Dans le cas de la voix, il est souvent utile d'avoir un contrôle indépendant de la fréquence fondamentale et des formants. Les méthodes appartenant à cette catégorie sont capables de modifier la fréquence fondamentale en respectant les positions et les caractéristiques des formants.

- **Le changement de hauteur simple.**

Celui-ci ne respecte pas les positions des formants, mais effectue une homothétie de l'axe des fréquences. Ce sont en général des méthodes moins spécifiques de la parole que les précédentes.

De nombreuses méthodes de modification du signal, reposant sur le principe (OLA : Over Lap and Add) de superposition/addition temporelle ont été proposées dans la littérature ces dernières années. [Hej90] [HeM91] [Kei03]

Parmi les plus importantes, citons les méthodes TDHS (Time Domain Harmonic Scaling), LSEE-MSTFTM , SOLA (Synchronized Overlap-Add) , Les modifications du signal permises par ces méthodes sont essentiellement des modifications de l'axe temporel du signal (compression/dilatation temporelle du signal). Pour cela, ces méthodes procèdent de manière tantôt proportionnelle tantôt synchrone à la période fondamentale, tantôt à l'analyse tantôt à la synthèse.

La méthode PSOLA (Pitch Synchronous Overlap-Add) se distingue de ces méthodes par une synchronie à la période fondamentale tant à l'analyse qu'à la synthèse. Ceci permet, à l'inverse des méthodes précédentes, un contrôle à la fois du déroulement de l'axe temporel et de la hauteur du signal.

La méthode de superposition/addition synchrone à la période fondamentale, PSOLA, repose sur la décomposition d'un signal s(t) en une série de formes d'onde élémentaires. Ces formes d'onde élémentaires sont obtenues par un fenêtrage exactement centré sur les périodes fondamentales du signal. Le signal de synthèse est alors reconstitué par superposition/addition (Overlap-Add) de ces formes d'onde élémentaires. La modification de la distance relative entre deux formes d'onde élémentaires, ainsi que la modification du nombre de formes d'onde élémentaires, permet de modifier la hauteur et l'axe temporel du signal.
Selon le nombre *n d*e périodes fondamentales renfermées par une forme d'onde élémentaire, nous parlerons de :

- méthode TD-PSOLA : n = 2
- méthode FD-PSOLA : n = 4

La décomposition du signal en formes d'onde élémentaires est effectuée par multiplication du signal $s(t)$ par une fenêtre de pondération $h(t)$ centrée en des temps m*i* appelés « marques de lecture » ou marques d'analyse. Ces marques de lecture sont positionnées de manière synchrone à la période fondamentale locale du signal.

2.2 La méthode TD-PSOLA

Cette méthode (dénommée TD-PSOLA, Time-Domain Pitch-Synchronous Overlap-Add) suppose que l'on traite un signal de parole dont on connaît la période.
L'idée [Mou90] est fondée sur l'hypothèse que le signal de parole est constitué d'impulsions glottales filtrées par le conduit vocal. On observe ainsi une succession de réponses impulsionnelles, positionnées à des temps multiples de la période (hypothèse du peigne temporel convolué avec la réponse impulsionnelle du conduit vocal).
On définit d'abord des « marques d'analyses » synchrones à la fréquence fondamentale pour les parties voisée, positionnées sur la forme d'onde à chaque période. Les modifications d'échelles sont alors effectuées de la façon suivante:

2.2.1 Modification de l'échelle temporelle.

Pour modifier la durée du signal sans en altérer la fréquence fondamentale, on va simplement dupliquer (étirement temporel) ou éliminer (compression temporelle) des périodes de la forme d'onde, en fonction du taux de modification désiré.

On est donc conduit à définir des marques de synthèse également synchrones au fondamental, associées aux marques d'analyse (de façon non-bijective puisque certaines marques sont dupliquées ou éliminées).

Les signaux à court-terme situés autour de chaque marque d'analyse sont alors extraits (par l'utilisation d'une fenêtre temporelle par exemple de type Hanning) de durée égale à deux périodes et centrée sur la marque d'analyse) et recopiés autour des marques de synthèse correspondantes, le signal modifié est alors obtenu par une simple méthode d' "overlap/add". La figure 2.1. illustre le principe de cette méthode pour un taux d'étirement temporel de 1.5.

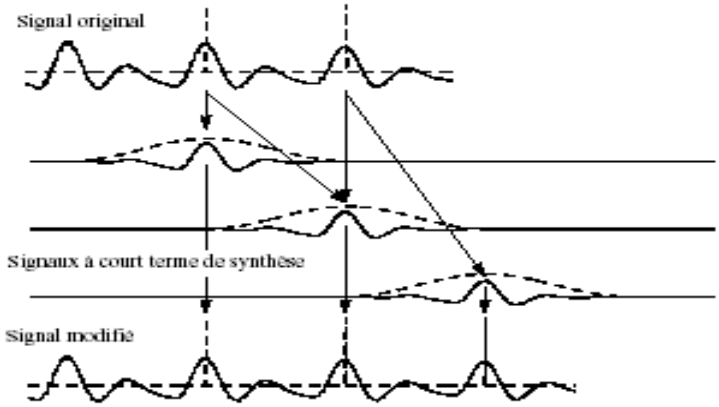

Figure 2.1 : Modification de la durée du signal par la méthode TD-PSOLA

On voit sur cette figure que deux périodes du signal original ont donné naissance à trois périodes dans le signal modifié, ce qui correspond bien à un étirement temporel mais la durée de la période n'est pas modifiée (l'écartement des marques de synthèse est le même que celui des marques d'analyse), la fréquence fondamentale du signal est conservée.

2.2.2 Modification de l'échelle fréquentielle.

Si l'on est capable de positionner dans le signal les marques d'analyse exactement sur le début de chaque onde glottale (réponse impulsionnelle du conduit vocal se produisant à chaque fermeture glottale), on conçoit que diminuer (resp. augmenter) l'intervalle de temps séparant deux marques d'analyse consécutives va permettre d'augmenter (resp. de diminuer) la fréquence du fondamental, sans que les formants soient modifiés (la réponse impulsionnelle n'est pas modifiée, en particulier sa décroissance temporelle et ses fréquences de résonance (les formants).

On est ainsi conduit à définir des marques de synthèse correspondant à la valeur modifiée du fondamental, et à les associer aux marques d'analyse comme précédemment. Puisque les marques de synthèse sont plus serrées (élévation du fondamental) ou écartées (abaissement du fondamental) que dans le signal original, il faut pour conserver la durée du signal dupliquer ou éliminer certaines marques. La figure 2.2. illustre le principe de cette méthode. On constate que les marques de synthèse étant plus écartées que les marques l'analyse, la période du signal est allongée. Pour éviter une élongation du signal, il est nécessaire d'éliminer périodiquement certains signaux à court-terme.

Lorsque le signal ne possède plus de fréquence fondamentale bien précise (cas des consonnes etc...), la modification est réalisée de façon non-synchrone, jusqu'à ce que l'on retrouve une région présentant un fondamental plus net.

La méthode décrite ci-dessus est appliquée principalement à la parole, et réalise des modifications de très bonne qualité. Par sa simplicité, elle peut faire l'objet d'une implémentation temps réel. En revanche, son application à des sons plus complexes, ou dénués de "pitch" (cas de la musique en général) pose de sérieux problèmes.

Les modifications de fondamental sont cependant très sensibles à la position des marques d'analyse.

Pour rendre la méthode plus robuste, les modifications de l'échelle fréquentielle peuvent être réalisées dans le domaine des fréquences (méthode FD-PSOLA) [MoC90].

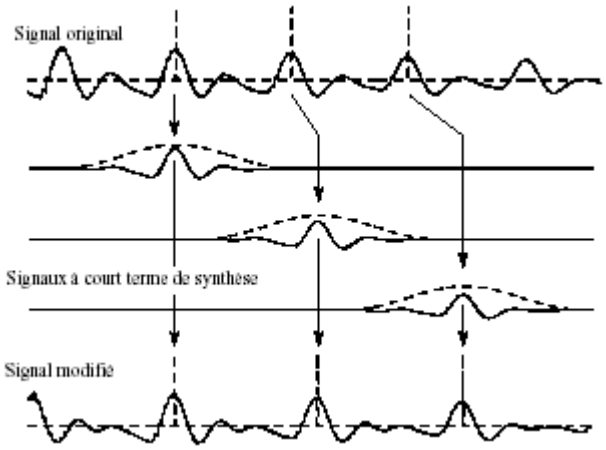

Figure 2.2 . Modification de la hauteur du signal par la méthode TD-PSOLA.

En haut, le signal original, au milieu trois signaux à court-terme générés à partir des trois premières marques d'analyse. En bas, signal modifié. L'écartement des marques de synthèse n'est pas identique à celui des marques d'analyse.

2.2.3 Modification conjointe temps et fréquence.

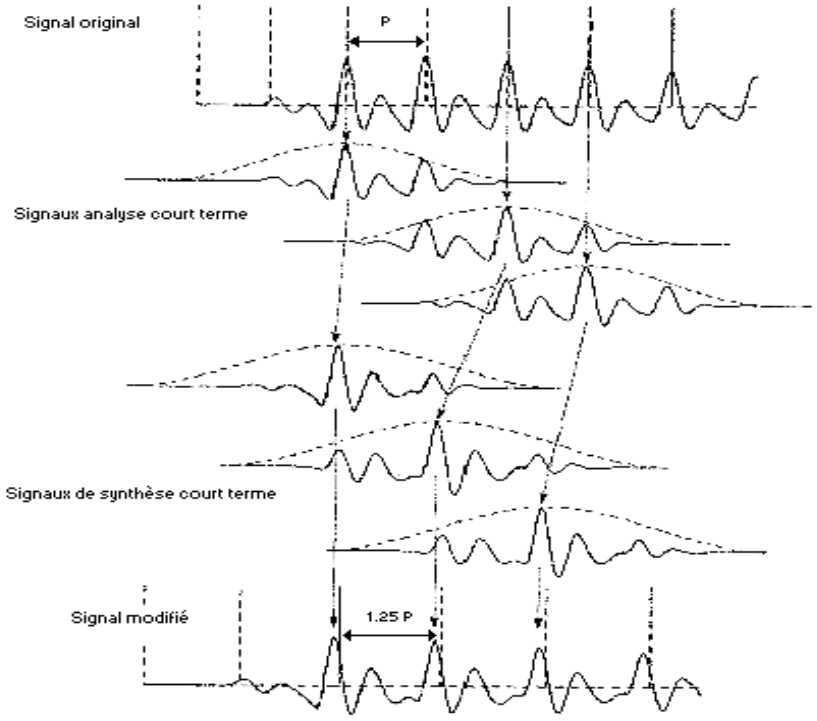

Figure 2.3: Modification conjointe de la durée et de la hauteur du signal par la méthode TD-PSOLA

2.3 La méthode FD-PSOLA

Dans la méthode FD-PSOLA, chaque forme d'onde élémentaire renferme quatre périodes fondamentales (n= 4).
Du fait de la durée de la fenêtre de découpage, le spectre de chaque forme d'onde élémentaire $s_i(t)$ présente une structure fine dans laquelle les harmoniques sont résolues.
Du fait de la résolution des harmoniques, la FD-PSOLA ne permet pas d'interprétation directe en termes d'approximation de l'enveloppe spectrale, ni en termes de décomposition source-filtre.

Une modification de hauteur du signal en FD-PSOLA ne peut plus se satisfaire d'une modification de la distance entre deux formes d'onde élémentaires successives. Ceci se comprend en constatant que le ré-échantillonnage du spectre aux multiples d'une fréquence $f \neq fo$ ne constitue plus un ré-échantillonnage de l'enveloppe spectrale mais un ré-échantillonnage du spectre harmonique.
La modification de hauteur en FD-PSOLA nécessite donc une correction conjointe du spectre d'amplitude et du spectre de phase des formes d'onde élémentaires.

2.4 Comparaison de TD-PSOLA et FD-PSOLA.

La méthode PSOLA est limitée au signaux non-mixtes (une seule source), monophoniques et périodiques.
De plus, la méthode TD-PSOLA, du fait de son fenêtrage étroit dans le domaine temporel, est limitée aux signaux dont la décroissance temporelle de la réponse impulsionnelle du filtre est rapide face à la période fondamentale. La méthode FD-PSOLA n'est pas limitée par cette contrainte du fait d'un fenêtrage temporel plus large.
Une manière de pallier cette contrainte de TD-PSOLA (qui pourrait également s'expliquer en terme de lissage des formants) a été proposée à travers l'utilisation jointe de TD-PSOLA et des méthodes de prédiction linéaire qui définit la méthode LP-PSOLA
La méthode LP-PSOLA consiste à appliquer le fenêtrage TD-PSOLA sur le signal résiduel (signal d'erreur de prédiction linéaire).
Une autre contrainte de TD-PSOLA est de positionner les marqueurs m*i* non seulement synchrones à la période fondamentale, mais également proches du maximum local

d'énergie à l'intérieur de chaque période. Seule la contrainte de périodicité est nécessitée par la méthode FD-PSOLA.

La méthode FD-PSOLA, ne donne pas lieu à une interprétation en termes de décomposition source/filtre. De ce fait, une modification de la hauteur du signal nécessite une correction des spectres des formes d'onde élémentaires. Cette correction introduit un certain nombre de nouveaux problèmes tels que la transposition des caractéristiques harmoniques/non-harmoniques, la transposition des détails fins du spectre, des relations de phase dans de nouvelles bandes de fréquence, ou l'apparition de trous dans le spectre. Des solutions ont cependant été proposées [MoC90] pour ces problèmes.
Un autre problème inhérent au fenêtrage large (n = 4) de FD-PSOLA est l'apparition d'une certaine réverbération dans le signal.

2.5 Traitement PSOLA sur le signal résiduel

Le traitement PSOLA sur le signal résiduel (résiduel de prédiction linéaire) a été proposé par Moulines dans [MoC90]. Le filtre de prédiction linéaire est calculé de manière synchrone à la période fondamentale (autour de chaque marque m*i*). Le signal résiduel est calculé par filtrage inverse et l'algorithme PSOLA appliqué au signal résiduel. Le traitement sur le signal résiduel (lorsque celui-ci peut être estimé) présente plusieurs avantages :

- Dans le cas de TD-PSOLA, la distorsion engendrée par le fenêtrage est moindre sur le signal résiduel que sur le signal. Ceci se comprend aisément en considérant la structure impulsionnelle du signal résiduel (Dans le cas du signal vocal, le signal résiduel constitue une approximation de la dérivée du signal glottal). Ceci se comprend également en considérant l'interprétation des formes d'onde élémentaires TD- PSOLA en termes d'estimation implicite de l'enveloppe spectrale. Dans le cas de LP-PSOLA, l'approximation de l'enveloppe spectrale est explicite et plus précise. Le traitement sur le signal résiduel permet également de simplifier le problème du positionnement des marqueurs m*i*, puisque l'influence des variations d'énergie locale du signal est minimisée par la soustraction de la contribution du filtre.

- Dans le cas de FD-PSOLA, un changement de hauteur nécessite une correction du spectre obtenue par un algorithme de compression/dilatation du spectre ou encore par un algorithme de répétition/élimination des harmoniques. Ces traitements ne permettent pas de respecter l'enveloppe spectrale du signal. L'application des algorithmes à un spectre à priori « blanchi » permet de conserver l'enveloppe spectrale.

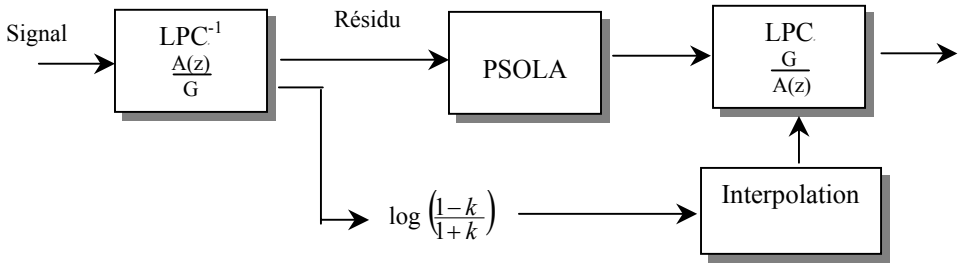

Fig. 2.4. Exemple de diagramme de la méthode LP-PSOLA (Interpolation Log Area)

Le diagramme de la méthode LP-PSOLA est illustré par la Fig.2.4

L'estimation des filtres est effectuée de manière synchrone à la période fondamentale autour des marques m*i*.

Le traitement PSOLA est effectué sur le signal résiduel ; le signal transformé est ensuite re-filtré. Ce filtrage s'effectue de manière synchrone autour des instants $\tilde{m}_j$. Les filtres sont obtenus par interpolation des coefficients log-area ratio, ou LSF de manière à garantir la stabilité des filtres.

2.6 Autres versions OLA .

Plusieurs recherches sur les algorithmes basées sur le principe OLA ont été menées pour améliorer la technique de base PSOLA.
Ainsi Peeters [Pee01] a introduit des modifications à la méthode PSOLA par l'introduction de la notion d'interpolation temporelle et fréquentielle aboutissant à deux nouvelles techniques appelées TDI-PSOLA et FDI-PSOLA. ainsi que la méthode SINOLA (contraction de Sinusoïdal et OLA) basée sur la le modèle de représentation sinusoïdale , et dont les grandes lignes sont décrites ci-dessous .
Lorsque la modification de l'axe temporel du signal ou de sa hauteur devient trop importante, l'algorithme de duplication/élimination introduit des discontinuités dans le signal :

- dans le cas de la duplication, par la reproduction d'une même forme d'onde puis passage à une autre,
- Dans le cas de l'élimination, par le passage d'une forme d'onde à une autre forme d'onde distante donc a priori non similaire.

Ceci se traduit perceptivement par une certaine rugosité du signal de synthèse.
Une solution face à cela est de créer de nouvelles formes d'onde élémentaires réduisant ces discontinuités par lissage. Ces formes d'onde élémentaires nouvelles sont obtenues par interpolation des formes d'onde associées aux marques m*i* et m*i*+1 voisines.
L'interpolation peut s'effectuer :

- soit dans le domaine temporel (Time Domain Interpolation PSOLA ou TDI-PSOLA) :
- soit dans le domaine fréquentiel (Frequency Domain Interpolation PSOLA ou FDI-PSOLA)

2.6.1 Interpolation temporelle des formes d'ondes élémentaires (TDI-PSOLA)

La forme d'onde élémentaire $\hat{s}_j(t)$ est obtenue par interpolation temporelle des formes d'onde élémentaires $s_i(t)$ TDI-PSOLA permet d'éviter les discontinuités de PSOLA standard.

Cependant, du fait d'une interpolation dans le domaine temporel de formes d'onde élémentaires proches mais non nécessairement alignées (soit du fait d'un mauvais positionnement des m*i*, soit du fait de l'erreur de troncature des m*i* vers leur équivalent en échantillon), l'algorithme TDI-PSOLA peut produire l'effet d'un filtrage passe-bas des formes d'onde élémentaires.

2.6.2 Interpolation fréquentielle des formes d'onde élémentaires (FDI-PSOLA)

La forme d'onde élémentaire $\tilde{s}_j(t)$ est obtenue par interpolation des spectres d'amplitude et de phase des formes d'ondes élémentaires $s_i(t)$.

La forme d'onde élémentaire interpolée est obtenue par TF inverse des spectres interpolés. Une attention particulière doit être portée afin d'éviter le repliement temporel du signal lors de la TF inverse.

Les spectres d'amplitude des formes d'onde élémentaires *si*(*t*) constituent une approximation de l'enveloppe spectrale du signal. L'enveloppe spectrale est supposée varier lentement sur la durée d'une période fondamentale. De ce fait, les spectres d'amplitude *Al* et *Al0* sont supposés proches (au sens d'une distance spectrale).

A l'inverse du spectre d'amplitude, le spectre de phase ne peut être considéré comme un signal à variation lente. Même dans le cas de deux formes d'onde élémentaires quasi-similaires, le positionnement des marques par rapport à chacune de ces formes d'onde élémentaire peut conduire à des spectres de phase très différents. Rajoutons à ce problème le fait que la phase n'est définie qu'en valeur principale *([-π;π])*. Une interpolation aveugle des spectres de phase peut conduire à la création d'une forme d'onde élémentaire $\tilde{s}_j(t)$ interpolée très différente de ses voisines, produisant une discontinuité dans le signal de synthèse, au lieu de l'effet de lissage recherché.

2.6.3 Modification du signal par la méthode SINOLA (SINusoidal OverLap-Add)

L'idée de cette méthode hybride, appelée SINOLA (SINusoidal OverLap-Add) est de combiner simultanément deux méthodes de modification du signal : la méthode PSOLA et la méthode basée sur le modèle sinusoïdal.

L'objectif a été de proposer une méthode de modification du signal utilisant simultanément les avantages de la méthode PSOLA et ceux du modèle sinusoïdal. La méthode hybride de modification du signal permettant le choix de celui le plus adapté aux caractéristiques locales du signal.

Notons enfin que de nombreuses méthodes de synthèse hybride ont été proposées. La plus connue est sans doute la méthode Harmonique plus Bruit [Sty96], étendue aux transitoires Harmonique plus Bruit plus Transitoires [Lev98].

2.7 Avantages de la méthode PSOLA.

La qualité des transformations du son obtenue par la méthode TD-PSOLA tient principalement à deux points :

1) le respect de l'enveloppe spectrale.
2) le respect des relations de phase entre les composantes fréquentielles lors d'une transformation de l'axe temporel (ceci tient au fait de l'absence de modification des formes d'onde et de la modification de l'axe temporel obtenu par superposition/addition).

La méthode PSOLA est particulièrement bien adaptée et robuste pour certaines classes de signaux sonores : les signaux harmoniques ou pseudo-harmoniques et présentant une forme d'onde dont la Réponse Impulsionnelle du filtre est courte par rapport à la période fondamentale du signal.

La méthode PSOLA présente l'avantage, par rapport à d'autres méthodes de transformation du son comme la synthèse sinusoïdale, de ne pas nécessiter l'estimation d'un modèle.

Même si la méthode PSOLA repose sur l'hypothèse d'un signal harmonique, le modèle harmonique n'est pas estimé. Ceci favorise non seulement la robustesse des

transformations (par exemple l'absence de bruit musical dans les hautes fréquences) mais également la préservation de caractéristiques difficilement modélisables comme les variations fines du signal au cours du temps et le jeu de relations de phase entre les composantes, la préservation de l'enveloppe spectrale, la possibilité de modifier les signaux pseudo-périodiques difficiles à modéliser en synthèse sinusoïdale.

De plus, son extension, permet également de l'utiliser pour les signaux non-voisés ou mixtes voisés/non-voisés.

2.8 Inconvénients de la méthode PSOLA.

Bien que la méthode PSOLA (TD-PSOLA) s'avère être un moyen simple et efficace pour la modification du pitch et de la durée des signaux de parole et du fait que la méthode procède dans le domaine temporel, l'enveloppe spectrale ne peut être contrôlée de manière efficace particulièrement pour un objectif de conversion de voix basé sur une transformation conjointe de l'enveloppe spectrale et de la fréquence fondamentale..

Chapitre 3

La conversion de voix

3. Introduction à la conversion de voix.

Dans la première partie de ce chapitre, nous exposons les principes de base d'un système de conversion de voix. Puis, dans la deuxième partie, nous dressons un état de l'art des techniques existant dans la littérature.

3.1 Principes d'un système de conversion de voix.

La plupart des systèmes de conversion de voix comportent dans leur élaboration deux phases principales. La première est une phase d'apprentissage ou de caractérisation durant laquelle des signaux de parole d'un locuteur source et ceux d'un locuteur cible sont analysés pour estimer une fonction de transformation.
La deuxième phase dite de transformation durant laquelle le système utilise la fonction de transformation précédemment apprise pour transformer de nouveaux signaux de parole source de façon qu'ils semblent, à l'écoute, avoir étés prononcés par le locuteur cible.

3.1.1 Phase d'apprentissage.

La figure 3.1 présente le principe général de la phase d'apprentissage d'un système de conversion de voix.

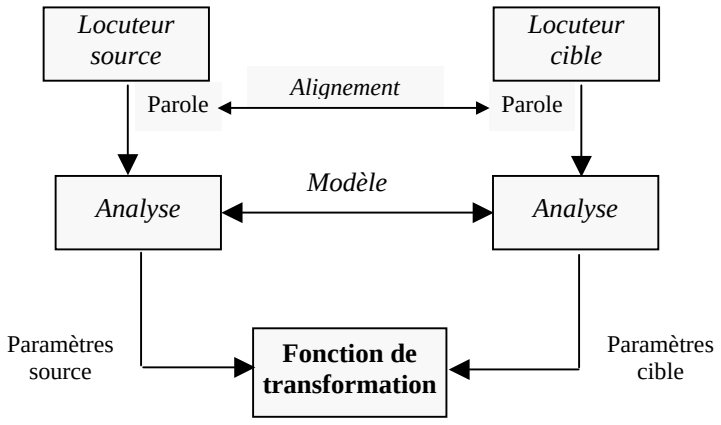

Figure 3.1 - Phase d'apprentissage en conversion de voix.

La phase d'apprentissage commence par une étape d'analyse des corpus de parole source et cible suivant le modèle mathématique adopté , on notera ici que les corpus doivent être parallèles c'est à dire qu'il comportent le même contenu phonétique. Cette phase d'analyse permet d'extraire les paramètres acoustiques utiles à l'estimation de la fonction de transformation.

La nature des paramètres acoustiques utilisés dépend du système de conversion. Généralement les travaux menés dans ce domaine sont basés sur la transformation de l'enveloppe spectrale, modélisée généralement par une variante des coefficients de prédiction linéaire (LPC, LSF, LAR), par cepstre discret ou par des paramètres relatifs aux formants ; auxquelles sont associées des modifications du pitch.[Kai01],[Gil03],[Enn05].

Par ailleurs et quels que soient les efforts fournis par les locuteurs source et cible (ou même dans le cadre d'un mimétisme). Les phrases émises par deux locuteurs doivent être préalablement alignées temporellement avant d'établir la fonction de transformation réalisant l'apprentissage.

L'alignement temporel est souvent réalisé à l'aide de l'algorithme DTW (voir Chap.4) dans les systèmes de conversion de voix présentés dans la littérature.[Gil03], [Tur03]. Cependant, il est possible de réaliser cet alignement avec d'autres méthodes comme les chaînes de Markov cachées ou par des approches connexionnistes.

3.1.2 - Phase de transformation

La stratégie commune à tous les systèmes de conversion de voix consiste à appliquer trame par trame la fonction de transformation à des signaux de parole source alors que le procédé global de transformation est décrit par la figure 3.2

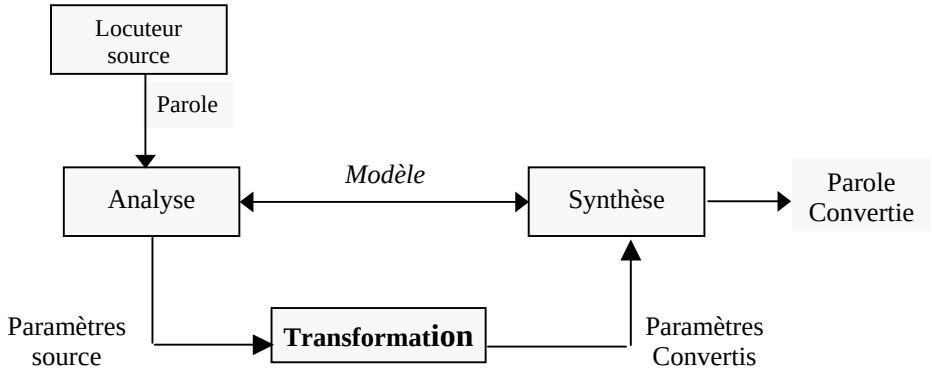

Figure 3.2 - Phase de transformation en conversion de voix

La procédure doit donc comporter une phase d'analyse permettant d'extraire les paramètres de la source pour chaque trame suivie d'une phase de conversion sur les paramètres à modifier par l'application de la fonctions de transformation, la phase finale étant la synthèse du signal de parole converti à base des paramètres modifiés et d'un éventuel résidu.

3.2 Fonctions de transformations

Les principales techniques de conversion présentées dans la littérature sont liées aux modélisations spectrales adoptées.[Enn05].

Deux modélisations de l'enveloppe spectrale sont souvent utilisées dans les fonctions de transformation en conversion de voix. Ce sont la modélisation AR et la modélisation cepstrale.
Le but recherché étant une paramétrisation utilisable en conversion de voix, la principale qualité requise est le fait de pouvoir supporter des transformations.
A ce titre, les coefficients du filtre LPC ne sont pas bien adaptes, car il est a priori difficile de contrôler l'effet d'une modification qui leur serait appliquée sur leur stabilité et sur la perception du signal produit. Les coefficients LSF (Line Spectral Frequency) semblent être ceux qui offrent les meilleures propriétés d'interpolation.
Pour plus de détails à ce sujet, des études comparatives ont été menées par Paliwal [PaK95] et [Pal95]ainsi que Grass dans [GrK90].

Les travaux précédents dans le domaine sont succinctement présentés dans l'ordre chronologique de leur réalisation.

1988 Abe et al. [Abe88]:

Les paramètres de filtre sont extraits et transformés par des techniques de quantification vectorielle (VQ) appliquées aux espaces acoustiques des locuteurs source et cible. Indépendamment de la méthode d'apprentissage utilisée la quantification vectorielle implique un regroupement très restrictif des données dans certaines régions de décision.

Son inconvénient principal est la limitation à un ensemble discret d'enveloppes, égal au nombre d'entrées dans un dictionnaire. De ce fait, il en résulte une qualité médiocre de la parole obtenue par cette conversion et la présence d'artefacts dus aux discontinuités spectrales.

1996 Stylianou et al. [Sty96]:

Selon son approche Stylianou a proposé l'utilisation des Modèles de Mélanges de Gaussiennes (GMM) afin de modéliser l'espace acoustique d'un locuteur.
Chaque classe est définie par son poids dans le mélange, sa moyenne et sa matrice de covariance.
Les paramètres du mélange de Gaussiennes du locuteur source sont estimés pour différents ordres par l'algorithme EM (Expectation Maximisation) puis une fonction de transformation est définie moyennant une distance quadratique moyenne entre les vecteurs transformés et les vecteurs cible.

Cette approche a donné naissance a plusieurs techniques dérivées et représente un standard en conversion de voix pour les travaux les plus récents.
Néanmoins son inconvénient majeur est dû au fait qu'elle est très gourmande en capacité mémoire et en termes de temps de calcul.

2001 Kain [Kai01]

Kain reprend le principe de l'utilisation des GMM mais en utilisant les LSF comme descripteurs acoustiques

2003 Gillett and King [Gil03]:

La transformation préconisée par Gillett comporte deux volets correspondant aux deux composantes du modèle source-filtre de production de la parole. Le premier volet transforme l'enveloppe spectrale représentée par le modèle de prédiction linéaire la transformation est réalisée par l'utilisation du modèle de mélanges de gaussiennes (GMM). Le second volet de cette approche réalise une prédiction des

détails spectraux sur les coefficients de prédiction linéaire transformés. Une nouvelle approche est proposée et est basée sur un dictionnaire de résidus.

2005 Ennejary [Enn05] :

Les approches étudiées dans cette thèse se basent sur des techniques de classification par GMM (Gaussian Mixture Model) et une modélisation du signal de parole par HNM (Harmonic plus Noise Model). Dans un premier temps, l'influence de la paramètrisation spectrale sur la performance de conversion de voix par GMM est analysée. Puis, la dépendance entre l'enveloppe spectrale et la fréquence fondamentale est mise en évidence. Deux méthodes de conversion exploitant cette dépendance sont alors proposées et évaluées.

3.3 Conversion de voix par la méthode Moyenne-Variance

3.3.1 Introduction au Contour de pitch

La méthode de transformation étant basée sur le principe de modification du contour de pitch nous avons jugé utile à ce niveau d'introduire quelques notions pour la bonne compréhension de la méthode.

Le contour de pitch se rapporte aux variations de la fréquence fondamentale f_0 au cours du temps. Puisque f_0 est défini seulement pour les sections voisées du signal de parole, le contour de pitch a des valeurs positives de pitch pour les groupes voisés séparés par des valeurs nulles pour des régions non voisées du signal de parole. Les régions non voisées correspondent à la condition où les cordes vocales ne vibrent pas, comme lors de la production de certaines plosives ou fricatives ainsi que pour les régions de silence. Le contour de pitch peut refléter les expressions émotives.

Par exemple, la plupart des phrases déclaratives ont un contour de pitch en déclinaison globalement.

D'autre part, les questions oui-non ont habituellement une reprise finale dans la dernière expression d'intonation qui fait parfois incliner le contour global. Tandis que de tels modèles sont plus ou moins indépendants du locuteur, il peut y avoir des tendances d'intonation spécifiques à un locuteur ou à un groupe de locuteurs.

La figure 3.3 illustre le contour de pitch pour la question « Did you buy any corduroy overalls ? » émise par un locuteur masculin britannique et un locuteur masculin américain. Les lignes de régression par les deux expressions indiquent clairement que le premier locuteur a un contour continûment relevé, tandis que le deuxième locuteur a autant d'élévations locales que de chutes ayant pour résultat une ligne de déclinaison globale. Quoique les deux lignes de régression aient une pente positive, qui est typique pour une question oui-non, la manière avec laquelle les locuteurs changent leurs contours sont tout à fait distincts.

Si de telles particularités sont observées entre paires de locuteurs, il serait souhaitable pour un bon système de conversion de pitch de réaliser un apprentissage de ces détails caractéristiques du locuteur.

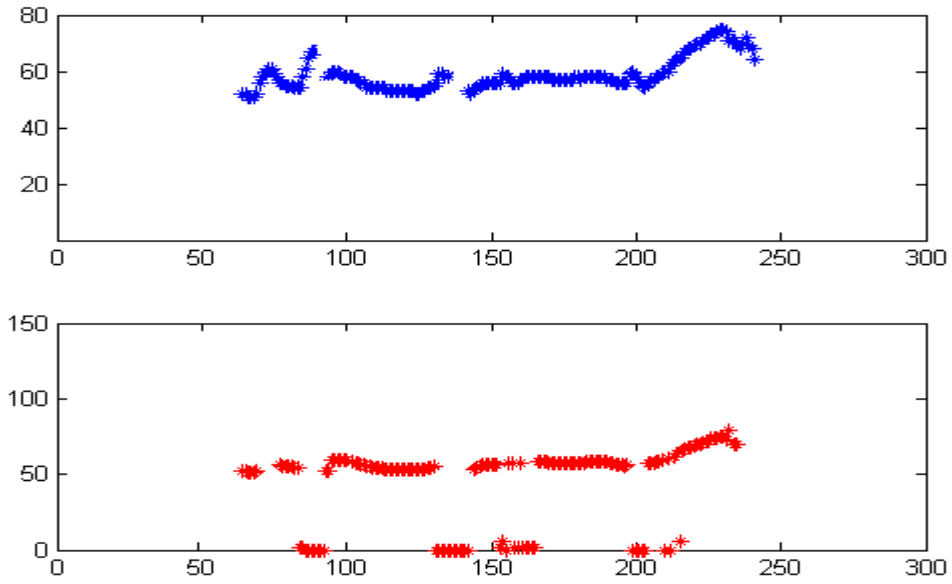

Figure 3.3 - Deux contours de pitch produits par deux locuteurs distincts

3.3.2 Transformation linéaire par Moyenne-Variance

Un travail substantiel a été entrepris sur la conversion spectrale à l'exemple de [KaM98] et [Ars99] impliquant un « mapping » ou correspondance entre les enveloppes spectrales de deux locuteurs au niveau segmental.

La transformation de pitch utilise pour ces auteurs la méthode linéaire de conversion « moyenne variance », avec la certitude que la fréquence moyenne de pitch véhicule la majeure partie de l'identité vocale du locuteur.

Cette méthode consiste à convertir les valeurs de pitch source f_0 tel que le contour converti soit assorti à la valeur moyenne de pitch et la gamme de pitch du locuteur cible, tout en maintenant le modèle d'intonation de la source. L'idée fondamentale est que les valeurs du pitch f_0 de chaque locuteur appartiennent à une distribution Gaussienne unique avec une moyenne et une variance spécifique.

Considérant le pitch cible comme une fonction du pitch source (fonction notée *h)*, une transformation linéaire peut alors être définie comme suit:

$$f_o^{cible} = h\,(f_o^{source}) = A.\,f_o^{source} + B \qquad \text{Eq. 3.1}$$

où f_o^{cible} est la valeur instantanée du pitch cible et f_o^{source} celle du locuteur source.

En réécrivant pour des raisons de commodité cette équation sous la forme :

$$c = h(s) = A.s + B \qquad \text{Eq.3.2}$$

L'objectif de la conversion revient à la détermination des paramètres A et B de la transformation linéaire, en termes de moyenne et variance des deux distributions Gaussiennes.

Si les valeurs de pitch cible ont une fonction de densité de probabilité notée f_c avec les paramètres correspondants (μ_c , σ_c) et les valeurs de pitch source une densité de probabilité notée f_s avec les paramètres (μ_s,σ_s) et en prenant l'inverse de la transformation linéaire de l'équation 3.2 on aura,

$$h^{-1}(c) = \frac{c - B}{A} \qquad \text{Eq. 3.3}$$

et tenant compte des propriétés de la fonction de densité de probabilité la transformation linéaire de l'équation 3.2 peut être réécrite comme suit :

$$f_c(c) = \frac{\partial h^{-1}(c)}{\partial c}\, f_s(h^{-1}(c)) \qquad \text{Eq.3.4}$$

De l'équation 3.3 on aura :

$$\frac{\partial h^{-1}(c)}{\partial c} = \frac{1}{A} \qquad \text{Eq. 3.5}$$

et l'équation 3.4 peut alors être réécrite comme suit :

$$f_c(c) = \frac{1}{A} f_s\left(\frac{c - B}{A}\right) \qquad \text{Eq. 3.6}$$

En remplaçant les deux membres de cette dernière égalité par leurs expressions pour une distribution Gaussienne avec moyenne et variance nous obtenons :

$$\frac{1}{\sqrt{2\pi\sigma_c^2}} \exp\left(-\frac{(c-\mu_c)^2}{2\sigma_c^2}\right) = \frac{1}{A}\frac{1}{\sqrt{2\pi\sigma_s^2}} \exp\left(-\frac{\left(\frac{c-B}{A}-\mu_s\right)^2}{2\sigma_s^2}\right) \qquad \text{Eq. 3.7}$$

En prenant le logarithme des deux termes de cette équation et en remplaçant dans l'équation 3.2 nous obtenons :

$$\log\left(\frac{A\sigma_s}{\sigma_c}\right) - \frac{(c-\mu_c)^2}{2\sigma_c^2} = -\left(-\frac{\left(\frac{c-B}{A}-\mu_s\right)^2}{2\sigma_s^2}\right) \qquad \text{Eq. 3.8}$$

les termes de second ordre peuvent alors être égalisés pour aboutir à une expression pour A en terme de variances des distributions de pitchs source et cible comme suit :

$$\frac{1}{2\sigma_c^2} = -\frac{1}{2A^2\sigma_s^2} \qquad \text{Eq. 3.9}$$

d'où :

$$A = \frac{\sigma_c}{\sigma_s} \qquad \text{Eq. 3.10}$$

En utilisant cette valeur pour l'équation 3.7 et en prenant la racine carrée des deux cotés on obtient :

$$\frac{-\left(\frac{\sigma_c S}{\sigma_s} + B - \mu_c\right)}{\sigma_c} = \frac{-(S - \mu_s)}{\sigma_s} \qquad \text{Eq. 3.11}$$

d'où

$$B = \mu_c - \mu_s \frac{\sigma_c}{\sigma_s} \qquad \text{Eq.3.12}$$

Un nombre fixe de phrases d'apprentissage peut être déterminé pour estimer les valeurs de moyenne et de variance pour chaque locuteur. Dans l'étape de conversion, l'équation 3.2 peut être appliquée à la valeur de f_0 de chaque trame voisée dans la phrase test pour produire un contour de cible.

Ce modèle est utilisé dans l'algorithme STASC (Speaker Transformation Algorithm using Segmental Codebooks) [Ars99] pour modéliser et transformer les caractéristiques du pitch.

Pour cela un facteur d'échelle de pitch variable dans le temps $\beta(t)$ est estimé en utilisant la valeur instantanée du pitch du signal à transformer. $\beta(t)$ est estimé en utilisant les valeurs statistiques des pitchs source et cible comme dans l'équation l'équation 3.12.

La conversion linéaire par la méthode moyenne-variance est une méthode couramment employée et réalise des résultats acceptables, néanmoins certains auteurs à l'exemple de [MoC90] ont constaté que les contours cible résultants ne possèdent pas généralement la structure prosodique fine du locuteur cible.

Chapitre 4

Outils de transformation

Dans ce chapitre nous avons regroupé quelques notions de base en traitement de la parole ainsi que les outils de traitement qui nous ont semblé avoir le plus d'importance par rapport au travail que nous présentons.

4.1 Alignement automatique

4.1.1 Introduction

Pour les besoins de traitements à partir de bases de données, nous avons souvent besoin de procéder à un alignement d'enregistrements issus de ces bases.
C'est pourquoi beaucoup de travail a été consacré aux méthodes automatiques et à la segmentation par l'alignement, pour la musique ainsi que pour les signaux de parole dans le cas de la reconnaissance de locuteurs par exemple.
Les deux principales méthodes utilisées sont l'alignement dynamique temporel (DTW Dynamique Time Warping) et le modèle de Markov caché (HMM pour Hidden Markov Model).

4.1.2 Applications

Plusieurs axes de recherche en informatique sont consacrés à l'automatisation des processus suivis par des humains. Par exemple, la segmentation d'une grande collection d'enregistrements, qui peuvent durer plusieurs heures, ne peut être faite manuellement en raison de la grande quantité de données. La même situation s'applique pour les signaux difficiles où la segmentation manuelle peut être pénible ou imprécise. Particulièrement pour les signaux musicaux.

- L'alignement permet l'indexation des médias continus avec une segmentation implicite, pour la récupération du contenu.
- Pour l'analyse musicale l'alignement permet une écoute à une position spécifique et l'extraction de paramètres expressifs d'interprétation du musicien
- Pour l'analyse du signal de parole l'alignement est lié au problème de la synchronisation en temps réel entre les locuteurs et la machine (ordinateur).

4.1.3 Dynamic Time Warping

L'alignement automatique qu'est la DTW (Dynamique TimeWarping dynamique) ou alignement dynamique temporel est une méthode largement employée.

Cette technique est basée sur la recherche du meilleur alignement global de deux séquences. Elle utilise un algorithme de recherche de chemin de Viterbi qui minimise les distances locales accumulées le long d'un chemin d'alignement pour réduire une distance globale entre les séquences.

Les séquences à aligner se composent de trames contenant des réalisations de manière générale. Les données des réalisations sont extraites par des techniques d'analyse de signal. Les données sont produites pour chaque trame selon un modèle de production. Dans notre cas, le modèle est celui de production de la parole.

L'alignement par DTW à la différence des méthodes basées sur HMM n'a pas besoin d'une étape d'apprentissage et de ce fait n'a pas besoin aussi de correction manuelle.
Notre choix pour cet algorithme est dû aussi à la possibilité d'optimisation de la place mémoire.
Les trames à aligner sont extraites par des techniques d'analyse de signal en utilisant la Transformée de Fourier à court terme.
La résolution temporelle de l'alignement dépend de la largeur de la fenêtre d'analyse ; par exemple, une fenêtre d'analyse de 256 points donne une résolution de 5,8 ms à une fréquence d'échantillonnage de 44,1 kHz.

Le nombre de trames à aligner est approximativement identique, de sorte que la diagonale soit le chemin idéal d'alignement. On notera cependant qu'il y a une imprécision inhérente à cela, car on ne connaît que la trame alignée et pas la position exacte dans la trame alignée. Habituellement, un choix arbitraire est fait en

prenant le milieu de la fenêtre pour la position précise (requise, par exemple, pour déterminer l'échantillon exact correspondant à une marque de synthèse de la parole).

4.1.3.1 Contraintes Locales.

Les distances locales et les contraintes locales et globales sont employées par l'algorithme de DTW pour calculer une matrice de distance augmentée adm(m,n) qui est le coût (la distance totale accumulée) du meilleur chemin jusqu'au point (m;n).

L'alignement est donné sous forme de chemin à travers la matrice de distance augmentée. Si un chemin passe par (m; n), la trame source m est alignée avec la trame cible. Voir exemple de la figure 4.1.

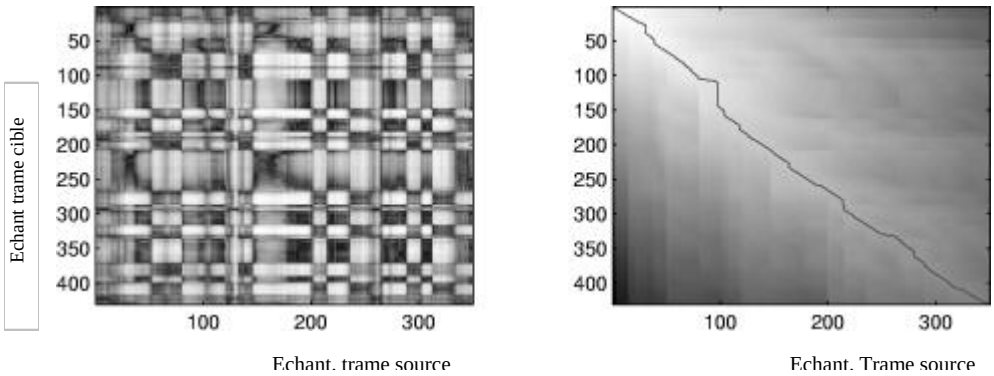

Figure. 4.1 Matrice de distance et chemin optimisé

Les trois différentes contraintes locales, également appelées voisinages, qui sont utilisées sont les contraintes de types I, III, et V, selon la terminologie utilisée par (Rabiner et Juang 1993). Le type I, le plus simple, donne déjà des résultats très acceptables par exemple pour l'élaboration de bases de données dans le domaine de la musique. [OrS01] [Sch98].

Cependant les types III et V sont plus « robustes » du point de vue performance d'alignement.

Les poids (pondérations) relatifs aux branches des contraintes locales sont notés ω_v pour la direction verticale, ω_h pour l'horizontale et ω_d pour la diagonale comme indiqué sur la figure 4.2 ils peuvent être accordés pour favoriser une direction de cheminement optimal.

Les contraintes locales de chemin définissent comment la matrice de distance au point (m; n) est calculée à partir des distances locales, avec la ldm(m; n) notée λ :

Type I:

$$adm\ (m,n)=\min\left\{\begin{array}{c} adm\ (m-1,n-1)+\omega_d\lambda \\ adm\ (m-1,n)+\omega_v\lambda \\ adm\ (m,n-1)+\omega_h\lambda \end{array}\right\} \qquad \text{Eq.4.1}$$

Type III:

$$adm\ (m,n)=\min\left\{\begin{array}{c} adm\ (m-1,n-1)+\omega_d\lambda \\ adm\ (m-2,n-1)+\omega_v\lambda \\ adm\ (m-1,n-2)+\omega_h\lambda \end{array}\right\} \qquad \text{Eq.4.2}$$

Type V :

$$adm\ (m,n)=\min\left\{\begin{array}{c} adm\ (m-1,n-1)+\omega_d\lambda \\ adm\ (m-2,n-1)+\omega_v\lambda+\omega_d ldm\ (m-1,n) \\ adm\ (m-1,n-2)+\omega_h\lambda+\omega_d ldm\ (m,n-1) \\ adm\ (m-3,n-1)+\omega_v\lambda+\omega_d ldm\ (m-2,n)+\omega_v ldm\ (m-1,n)) \\ adm\ (m-1,n-3)+\omega_h\lambda+\omega_d ldm\ (m,n-2)+\omega_h ldm\ (m,n-1)) \end{array}\right\}$$

Eq.4.3

La contrainte type I est la seule à ne permettre que les directions horizontale et verticale dans le choix du chemin ; l'inconvénient de ce type de contrainte est que le chemin peut se « coincer dans une vallée… »induisant ainsi une distance locale erronée.

Les types III et V contraignent la pente à être respectivement entre 2 et 1/2 ou 3 et 1/3 i.e. que le chemin est forcé dans les deux directions d'où leur meilleure performance.

Les valeurs standards pour les poids de contrainte local ω_v, ω_h, ω_d = (1, 1, 2) pour le type I et V ou (3, 3, 2) pour le type III, ne favorisent aucune direction et donnent de bons résultats.

L'expérience montre aussi que de faibles valeurs de ω_d favorisent la diagonale.

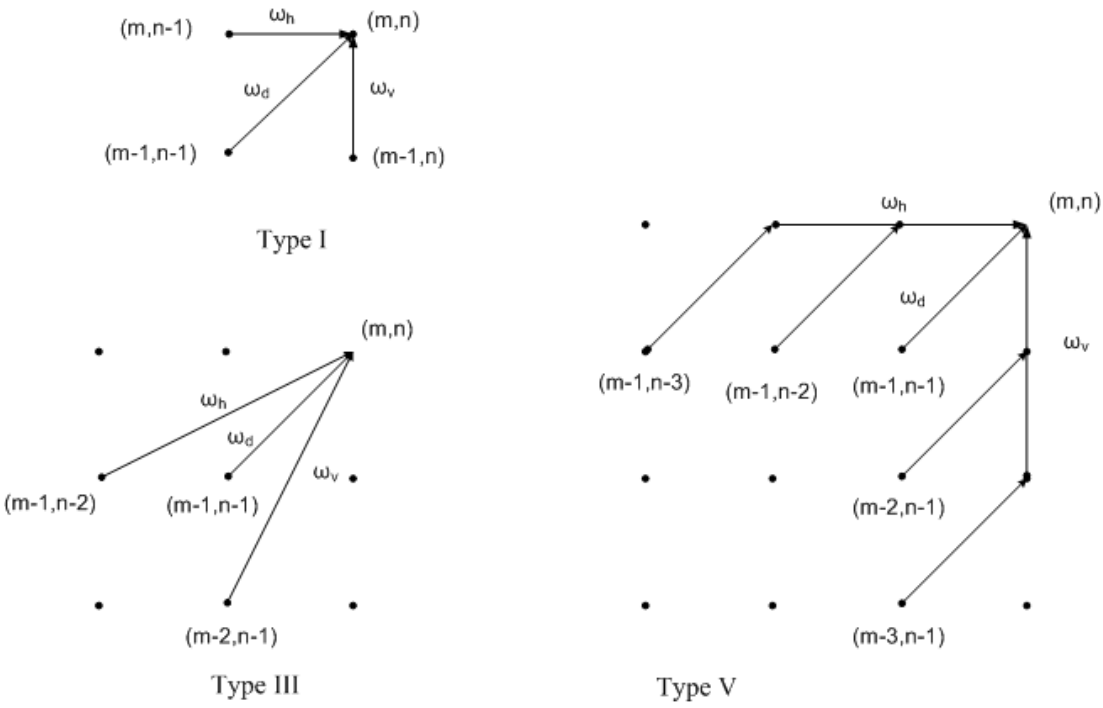

Figure 4.2 Voisinage du point (m,n) pour les types I, II et V

4.1.3.2 L'algorithme DTW

L'alignement DTW est basé sur l'algorithme de Viterbi .On a vu que l'algorithme de DTW recherche le meilleur chemin d'alignement entre deux réalisations selon des distances locales et un certain nombre de contraintes locales et globales.

Les distances locales sont stockées dans une matrice de distances locales mdl(m,n) où chaque valeur mdl(m;n) exprime le décalage entre la trame cible et la trame source m.

Ensuite, un chemin est calculé pour minimiser la distance ajoutée entre les points (1, 1) et (M,N) avec M et N respectivement nombre de fragments de la source et de la cible.

Il est calculé de façon itérative, en mettant à jour la matrice des distances ajoutées mda(m, n), qui donne la distance du meilleur chemin reliant (1, 1) de (m, n). Une matrice, ψ(m, n) conserve les coordonnées du point précédant dans le chemin. Il existe différents types de méthodes DTW, mais toutes ne différent que dans le calcul de mda(m, n).

En plus des contraintes locales définies plus haut, les contraintes globales suivantes doivent être considérées: le point extrêmes de l'alignement doivent être fixés à (1,1) et (M,N) où M et N sont le nombre de trames source et cible respectivement.

Le chemin optimal devrait alors être près de la diagonale, de façon à éviter sa déviation.

4.1.3.3 Résultats de l'alignement de DTW

Nous présentons dans ce qui suit un exemple d'alignement par DTW sur des signaux de parole. Les phrases sont issues de la base TIMIT (voir Annexe 1) .

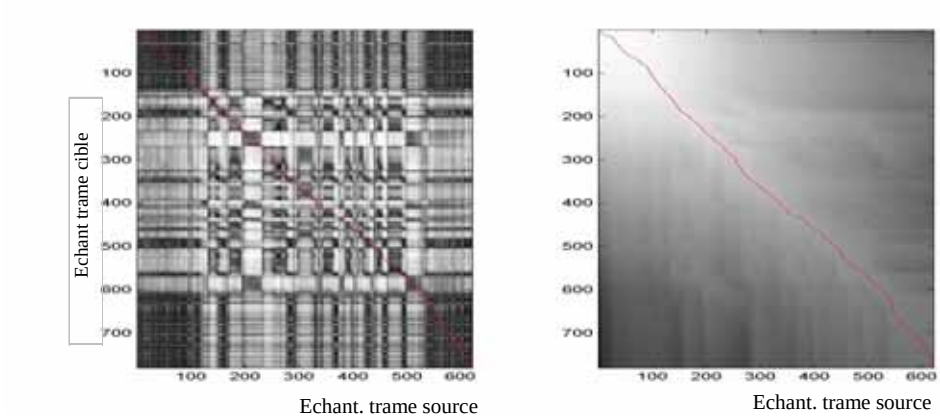

Figure. 4.3 Matrice de distance et chemin optimisé

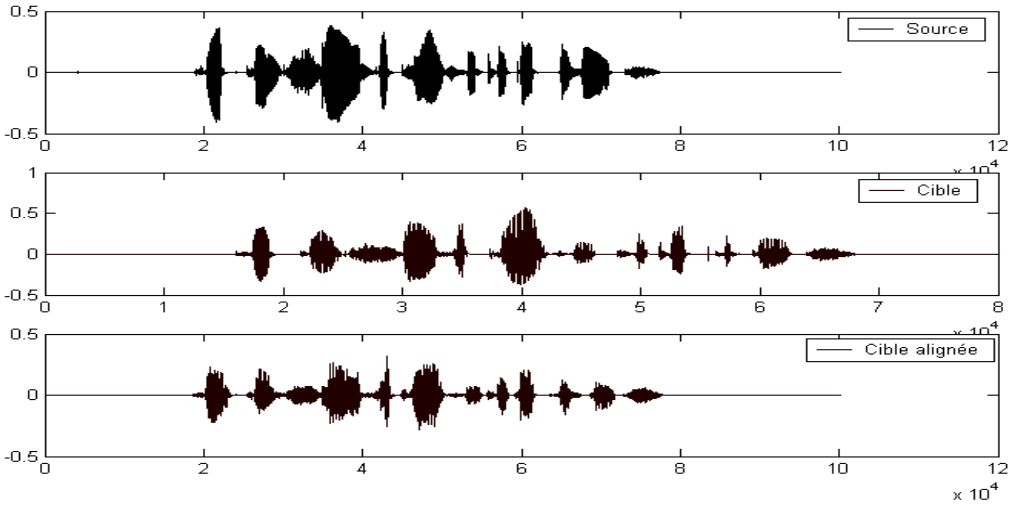

Figure. 4.4 Phrases alignées par DTW

4.2 Modélisation par prédiction linéaire

4.2.1 Introduction

L'analyse de la parole est une étape devenue indispensable à toute application de synthèse, de codage ou de reconnaissance dans le domaine. Elle repose en général sur un modèle.

L'analyse consiste donc en l'estimation des paramètres du modèle dans le but de lui faire correspondre le signal analysé.

S'il existe de nombreux modèles de parole, il en est un que l'on retrouve très souvent, le modèle prédictif linéaire(***LPC*** *linear prédictive coding*) ; un outil fréquemment utilisé qui a suscité de nombreuses études depuis une trentaine d'années. [Mar72] [Mak75]…

De la même façon qu'un signal de parole réel est produit par le passage, à travers le filtre que constitue notre conduit vocal, d'un signal d'excitation crée par les poumons et les cordes vocales, ce même signal de parole peut être modélisé par le passage d'un signal d'excitation numérique à travers un filtre numérique récursif. Le signal d'excitation sera tantôt une suite d'impulsions numériques (qui serviront à simuler les impulsions de débit crées par les cordes vocales), il s'agit alors d'un son *voisé*, tantôt du bruit numérique (qui reproduira le souffle poussé par les poumons), c'est un son *non voisé.* Ce modèle est appelé prédictif linéaire en raison du fait qu'il correspond à une régression linéaire très simple entre le signal d'excitation et le signal vocal produit.

Particulièrement adapté à la modélisation d'enveloppe spectrale des signaux de parole, nous en étudions dans ce chapitre les principes théoriques avant de nous intéresser à ses avantages et inconvénients dans le cadre de la conversion de voix.

4.2.2 Principe de la prédiction linéaire

4.2.2.1. Introduction

Sur le système d'entrée/sortie modélisé Figure 4.5, le signal de sortie *Sn* s'écrit comme une combinaison linéaire des échantillons du signal de sortie observés aux P instants précédents, et des échantillons du signal d'entrée U_n observés à l'instant présent et aux Q instants précédents (Modèle ARMA, Auto Regressive

Moving Average).

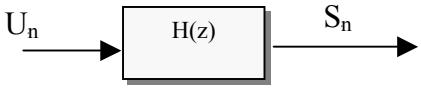

Figure 4.5 – Système d'entrée /sortie

$$S_n = \sum_{k=1}^{P} a_k S_{n-k} + G \sum_{l=0}^{Q} b_l U_{n-l} \qquad , b_0 = 1 \qquad \text{Eq.(4.8)}$$

où G et le couple $\{a_k\},\{b_l\}$ représentent respectivement le gain et les coefficients du filtre H.

Dans le domaine fréquentiel, en désignant par H(z) la fonction de transfert du système, l'équation (4.8)
s'écrit :

$$H(z)=\frac{S(z)}{U(z)}=G.\frac{1+\sum_{l=1}^{Q} b_l.z^{-l}}{1-\sum_{k=1}^{P} a_k.z^{-k}} \qquad \text{Eq. (4.9)}$$

Où

$$S(z)=\sum_{n=-\infty}^{\infty} S_n.z^{-n} \qquad \text{Eq. (4.10)}$$

Les racines du numérateur et du dénominateur sont respectivement les zéros et les pôles du modèle.

Dans le cas où $b_l = 0$, quel que soit $1 \leq l \leq Q$, *H(z)* se réduit à un modèle tout-pôles, appelé modèle autorégressif. L'équation (4.8) devient dans ce cas:

$$S_n=\sum_{k=1}^{P} a_k.S_{n-k}+G.U_n \qquad \text{Eq. (4.11)}$$

4.2.2.2. Modèle autorégressif

On tente d'approcher S_n avec les échantillons observés aux instants précédents.
Considérant la prédiction linéaire d'ordre P du signal S_n, le signal de prédiction, à l'instant n, s'écrit comme une combinaison linéaire des P échantillons passés :

$$\widetilde{S}_n = \sum_{k=1}^{P} a_k . S_{n-k} \qquad \text{Eq.(4.12)}$$

Les coefficients $\{a_k\}$ sont appelés coefficients de prédiction.
La différence entre le signal S_n et sa valeur prédite $\widetilde{S}_n$ constitue l'erreur de prédiction ou résidu :

$$e_n = GU_n = S_n - \widetilde{S}_n = S_n - \sum_{k=1}^{P} a_k . S_{n-k} \qquad \text{Eq.(4.13)}$$

Le filtre, décrit par l'équation (4.13), est dit blanchisseur car l'opération de prédiction linéaire a pour conséquence de décorréler les valeurs de l'erreur de prédiction [Mak75]. Le filtre blanchisseur a pour transformée en z :

$$A(z) = 1 - a_1 z^{-1} + \ldots + a_p z^{-p} = 1 - \sum_{k=1}^{P} a_k . z^{-k} \qquad \text{Eq. (4.14)}$$

En appliquant sur le résidu e_n le filtre inverse de fonction de transfert $\frac{1}{A(z)}$,on obtient le signal *Sn*.

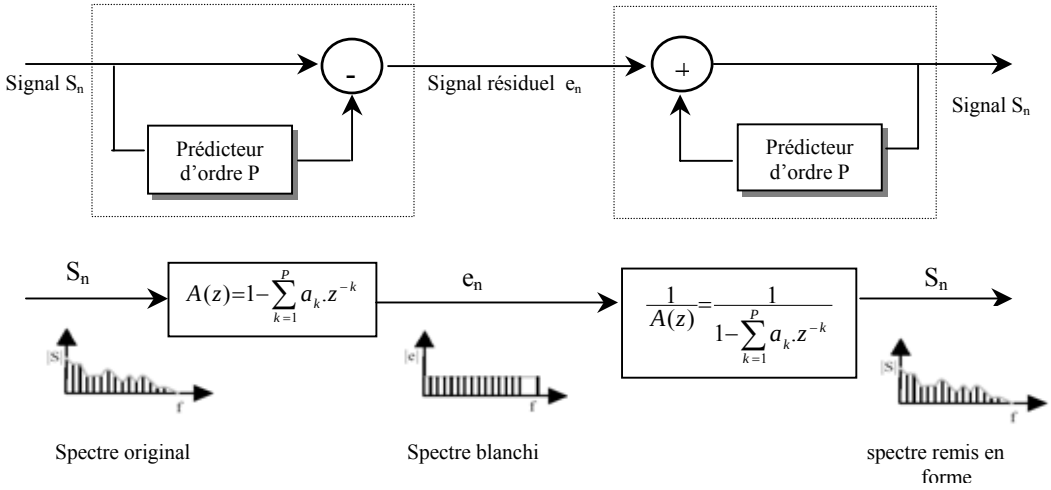

Figure 4.6 Filtre blanchisseur et filtre de synthèse

4.2.2.3. Calcul des coefficients de prédiction

On distingue classiquement deux méthodes pour estimer les coefficients de prédiction :

- La méthode de l'autocorrelation
- La méthode de la covariance

- **Méthode de l'autocorrélation**

Dans cette méthode, le signal S_n est multiplié par une fenêtre w_n .On obtient ainsi le signal fenêtré :

$$S'_n = w_n S_n \qquad \text{Eq. (4.15)}$$

On minimise ensuite l'énergie du signal d'erreur E défini par :

$$E = \sum_{n=-\infty}^{\infty} e_n^2 = \sum_{n=-\infty}^{\infty} (S'_n - \sum_{k=1}^{P} a_k - S'_{n-k})^2 \qquad \text{Eq. (4.16)}$$

La recherche des coefficients $\{a_k\}$ se fait en minimisant E relativement aux coefficients $a_1 \ldots a_k$

En dérivant *E* par rapport aux coefficients $\{a_k\}$

$$\frac{\partial E}{\partial a_k}=0\,,\ k=1\,,\ldots\ p \qquad \text{Eq.(4.17)}$$

On obtient les p équations suivantes :

$$\sum_{k=1}^{P} a_k . \sum_{n=-\infty}^{\infty} S'_{n-i} S'_{n-k} = \sum_{n=-\infty}^{\infty} S'_n S'_{n-i} \quad ,\ 1 \le i \le p \qquad \text{Eq.(4.18)}$$

Dans les équations (4.18), on ne considère que les données sont nulles à l'extérieur de la fenêtre d'analyse

En définissant la fonction d'autocorrelation du signal fenêtré S_n' :

$$R(i)=\sum_{n=-\infty}^{\infty} S'_n S'_{n-i} = \sum_{n=i}^{N-1} S'_n S'_{n-i} \quad ,\ 1 \le i \le p \qquad \text{Eq.(4.19)}$$

où *N* représente la longueur de la fenêtre d'analyse, et en substituant les équations (4.19) aux équations (4.18), on obtient les équations suivantes :

$$\sum_{k=1}^{P} a_k R(i-k)=R(i) \qquad ,1 \le i \le p \qquad \text{Eq. (4.20)}$$

Cette dernière équation peut s'écrire sous la forme matricielle :

$$\begin{bmatrix} R_0 & R_1 & \ldots R_{p-1} \\ R_1 & R_0 & \ldots R_{p-2} \\ \ldots & \ldots & \ldots \ \ldots \\ R_{p-1} R_{p-2} \ldots & & R_0 \end{bmatrix} \begin{bmatrix} a_1 \\ a_2 \\ \ldots \\ a_p \end{bmatrix} = \begin{bmatrix} R_1 \\ R_2 \\ \ldots \\ R_p \end{bmatrix} \qquad \text{Eq. (4.21)}$$

Ce système matriciel se résout en tenant compte du fait que la matrice d'autocorrélation est une matrice de Toeplitz. Cette propriété permet de résoudre efficacement, c'est-à-dire sans inversion de la matrice *R(i)*, le système par l'algorithme de Levinson-Durbin.

Cette propriété assure également que le filtre *A(z)* est à phase minimale.

Dans le filtre de synthèse, $H(z)=\frac{1}{A(z)}$ les zéros de *A(z)* deviennent les pôles de *H(z)* et le fait que *A(z)* soit à phase minimale garantit la stabilité du filtre de synthèse *H(z)*.

- **Méthode de la covariance**

Dans la méthode de la covariance, on fenêtre le signal d'erreur (au contraire de la méthode de l'autocorrélation dans laquelle on fenêtre le signal S_n).
L'énergie E du signal s'écrit alors :

$$E = \sum_{n=-\infty}^{\infty} (e'_n)^2 = \sum_{n=-\infty}^{\infty} (e_n^2) w_n \qquad \text{Eq .(4.22)}$$

En dérivant E par rapport aux coefficients {ak}, on obtient les p équations linéaires suivantes

$$\sum_{k=1}^{P} \Phi(i,k).a_k = \Phi(i,0) \qquad 1 \le i \le p \qquad \text{Eq. (4.23)}$$

où $\Phi(i,k)$ est la fonction de covariance du signal définie par :

$$\Phi(i,k) = \sum_{n=-\infty}^{\infty} w_n S_{n-i} S_{n-k} \qquad \text{Eq. (4.24)}$$

Les équations (4.23) peuvent s'écrire sous la forme matricielle suivante :

$$\begin{bmatrix} \Phi(1,1)\Phi(1,2)...\Phi(1,P) \\ \Phi(2,1)\Phi(2,2)...\Phi(2,P) \\ ... \quad ... \quad .. \quad ... \\ \Phi(P,1)\Phi(P,2)..\Phi(P,P) \end{bmatrix} \begin{bmatrix} a_1 \\ a_2 \\ ... \\ a_P \end{bmatrix} = \begin{bmatrix} \Phi(1,0) \\ \Phi(2,0) \\ ... \\ \Phi(P,0) \end{bmatrix} \qquad \text{Eq (4.25)}$$

Où . $\Phi(i) = \Phi(i,0)$, $i=1,2,\ldots,p$

Cette matrice est symétrique mais les coefficients sur les diagonales ne sont pas égaux entre eux, à la différence de la matrice d'autocorrélation définie ci-dessus. La méthode de décomposition de Cholesky permet de résoudre ce système.

L'intérêt de la méthode de covariance est de ne faire aucune hypothèse concernant les données à l'extérieur de l'intervalle d'analyse, offrant ainsi une estimation spectrale plus fine. Cette méthode permet d'estimer plus précisément l'enveloppe spectrale, et de conserver une bonne précision temporelle car elle minimise l'erreur sur un intervalle fini. L'inconvénient est que contrairement à la méthode de l'autocorrélation, la stabilité du filtre tout-pôle n'est pas assurée.

4.2.3 Représentation des coefficients de prédiction

Les coefficients {ak}, ou coefficients LPC (Linear Predictive Coding), offrent des caractéristiques qui ne facilitent pas le codage. Une faible erreur de quantification de l'un de ces paramètres entraîne de fortes variations dans le spectre restitué par l'ensemble du filtre et génère souvent des problèmes d'instabilité au filtre de synthèse. Les paramètres LSF (Line Spectral Frequencies), appelés encore paramètres LSP (Line Spectral Pair), possèdent en revanche des propriétés de quantification plus appropriées et permettent un meilleur contrôle de la stabilité.
Notons que la conversion est réversible et sans perte. Elle est fournie en annexe B.

Le modèle source-filtre associé aux signaux de parole, illustré Figure 4.7, est particulièrement adapté à la représentation paramétrique de ces sons.

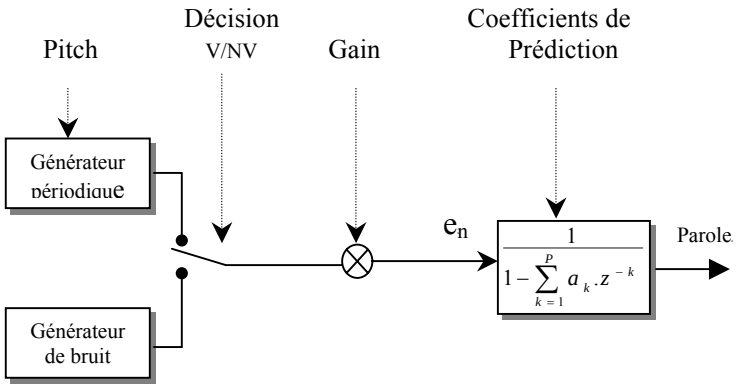

Figure 4.7 : Modèle paramétrique de production de la parole

- Les signaux de parole voisée peuvent être modélisés efficacement par un peigne harmonique mis en forme par une enveloppe spectrale adéquate.
- Les signaux de parole non-voisée peuvent être modélisés efficacement par du bruit mis en forme par une enveloppe spectrale de bruit associée.

4.2.3.1 Transformation des coefficients LSP (Line Spectral Pair)

Les coefficients LSP sont liés à la position des pôles sur l'axe des fréquences et ont la propriété d'être rangés par ordre croissant; cet agencement des paramètres permet de prendre en compte des critères perceptifs et offre une propriété de codage ou de transformation efficace.

4.2.3.2. Lissage des coefficients LSP

De faibles variations de l'enveloppe spectrale entre deux trames consécutives peuvent entraîner une modification importante des coefficients LPC lors de l'analyse. A la synthèse, ces faibles variations peuvent engendrer des discontinuités temporelles lors de la synthèse du signal.
L'interpolation des filtres permet de résoudre ces problèmes de discontinuité. La qualité de restitution des signaux s'en trouve ainsi fortement améliorée sans exiger d'information additionnelle.
La technique consiste à interpoler linéairement les coefficients LSP calculés sur une trame de durée T de façon à les appliquer, à la synthèse, sur des sous-trames de durée plus faible. Ainsi pour deux trames d'analyse consécutives de 20ms, on interpolera avantageusement les coefficients LSP afin d'appliquer les filtres de synthèse correspondants sur des trames de durée plus courte (typiquement de T/8 = 2.5 ms).

Une fois les coefficients LSP transmis au décodeur, ils sont convertis en coefficients LPC (méthode de conversion décrite en annexe B). L'ajustement d'enveloppe est réalisé par filtrage du signal d'excitation (signal résiduel *en* spectralement blanchi décrit Figure 4.6) par le filtre de synthèse LPC déduit de ces coefficients, et de transformée en z :

$$\frac{1}{A(z)} = \frac{1}{1 - \sum_{k=1}^{P} a_k z^{-k}} \qquad \text{Eq(4.26)}$$

4.2.3.3 Conversion LPC/LSP *

A partir du polynôme d'ordre P ([UIT 96])

$$A(z)=1-\sum_{k=1}^{P} a_k z^{-k} \qquad \text{Eq.(4.27)}$$

on construit les deux polynômes réciproques d'ordre P+1

$$\begin{aligned} F_1'(z) &= A(z)+z^{-(P+1)}A(z^{-1}) \\ F_2'(z) &= A(z)-z^{-(P+1)}A(z^{-1}) \end{aligned} \qquad \text{Eq.(4.28)}$$

Ces deux polynômes ont les propriétés suivantes :

- $F_1'(z)$ est symétrique et $F_2'(z)$ est antisymétrique (conjugués)
- Si toutes les racines de A(z) sont à l'intérieur du cercle unité alors les racines de $F_1'(z)$ et de $F_2'(z)$ sont sur le cercle unité
- Les racines de $F_1'(z)$ et de $F_2'(z)$ apparaissent de façon alternée sur le cercle unité.

Les polynômes $F_1'(z)$ et $F_2'(z)$ possèdent respectivement les racines z = -1 ($\omega=\pi$) et z = +1 ($\omega=0$). On élimine ces deux racines en définissant les deux nouveaux polynômes suivants :

$$\begin{aligned} F_1(z) &= \frac{F_1'(z)}{1+z^{-1}} \\ F_2(z) &= \frac{F_2'(z)}{1-z^{-1}} \end{aligned} \qquad \text{Eq.(4.29)}$$

Chacun de ces polynômes possède P/2 racines conjuguées sur le cercle unité et on peut les écrire comme suit :

Note (*) : Selon la recommandation UIT-T G729, Mars 1996 de l'Union internationale des télécommunications,

$$F_1(z)=\prod_{i=1,3,\ldots,(P-1)}(1-2.q_i.z^{-1}+z^{-2})$$
$$F_2(z)=\prod_{i=2,4,\ldots,(P)}(1-2.q_i.z^{-1}+z^{-2}) \qquad \text{Eq.(4.30)}$$

avec qi = cos(ωi).
Les coefficients LSF correspondent à la position angulaire ωi, entre 0 et π, des racines de P(z) et de Q(z) et vérifient :

$$\omega_0 = 0 < \omega_1 < \ldots < \omega_p < \omega_{p+1} = \pi \qquad \text{Eq (4.31)}$$

Etant donné que les deux polynômes F1(z) et F2(z) sont symétriques, il suffit de calculer les P/2 premiers coefficients de chaque polynôme, au moyen des relations de récurrence suivantes :

$$f1(i+1) = a_{i+1} + a_{10-i} - f1(i) \qquad i = 0,\ldots,P/2-1$$
$$f2(i+1) = a_{i+1} + a_{10-i} - f2(i) \qquad i = 0,\ldots,P/2-1 \qquad \text{Eq.(4.32)}$$

où f1(0) = f2(0) = 1

Les coefficients LSF sont trouvés par évaluations des polynômes F1(z) et F2(z) à 60 points équidistants entre 0 et π puis en vérifiant les changements de signes. Chacun de ces derniers correspond à l'existence d'une racine et l'intervalle du changement de signe est alors divisé par 4 afin d'affiner la recherche de la racine.

On se sert des polynômes de Chebycheff pour évaluer les deux polynômes $F_1(z)$ et $F_2(z)$. Grâce à cette méthode, les racines sont calculées directement dans le domaine des cosinus. Le polynôme $F_1(z)$ ou $F_2(z)$, évalué à la valeur $z = e^{j\omega}$, peut s'écrire comme suit:

$$F(\omega) = 2\, e^{-j\omega P/2}.C(x) \qquad \text{Eq.(4.33)}$$

où C(x) = TP/2(x) + f(1)TP/2-1(x) + …+ f(P/2-1)T1(x) + f(P/2)/2

et où le terme $Tm(x) = \cos(m\omega)$

est le polynôme de Chebycheff du nième ordre et où les f(i), i=1,…,P/2, sont les coefficients du polynôme F1(z) ou F2(z), calculés par l'équation (10). Le polynôme C(x) est évalué à une certaine valeur de $x = \cos(\omega)$ au moyen de la relation de récurrence suivante :

Pour k = P/2-1 à 1

$$B_k = 2xb_{k+1} - b_{k+2} + f(P/2-k) \qquad \text{Eq. (4.34)}$$

Fin

$$C(k) = xb_1 - b_2 + f(P/2)/2$$

Avec les valeurs initiales $b_{P/2} = 1$ et $b_{P/2+1} = 0$.

4.2.3.4 Conversion LSP-LPC

Une fois quantifiés, les coefficients LSF sont reconvertis en coefficients LPC, ai.

Pour ce faire, on détermine les coefficients des polynômes F1(z) et F2(z) par expansion des équations (4.9), sur la base des coefficients LSF quantifiés. A partir des coefficients qi, on calcule les coefficients f1(i), i=1,…,P/2, par la relation de récurrence suivante :

Pour i=1 à P/2

$$f_1(i) = -2.q_{i-1}.f_1(i-1) + 2f_1(i-2)$$

Pour j = de i-1 à 1

$$f_1(j) = f_1(j) - 2q_{2i}-1.f_1(j-1) + f_1(j-2) \qquad \text{Eq(4.35)}$$

avec comme valeurs initiales $f_1(0) = 1$ et $f_1(-1) = 0$. Les coefficients $f_2(i)$ sont calculés de la même manière, avec remplacement des q_{2i-1} par q_{2i}.

Une fois les coefficients $f_1(i)$ et $f_2(i)$ calculés, on multiplie les polynômes $F_1(z)$ et $F_2(z)$ par $(1+z^{-1})$ et $(1-z^{-1})$ respectivement, pour obtenir les polynômes $F_1'(z)$ et $F_2'(z)$, qui donnent les coefficients suivants :

$$f1'(i) = f_1(i) + f_1(i-1) \quad , \quad i = 1,\ldots,P/2$$
$$f2'(i) = f2(i) - f2(i-1) \quad , \quad i = 1,\ldots,P/2 \qquad \text{Eq (4.36)}$$

Finalement, on retrouve les coefficients LPC à partir des coefficients $f_1'(i)$ et $f_2'(i)$ comme suit :

$$a_i = \begin{cases} 0{,}5.f_1'(i)+0{,}5.f_2'(i) \\ \qquad\qquad\qquad\qquad i= 1,\ldots, P/2 \\ 0{,}5.f_1'(11-i)+0{,}5.f_2'(11-i) \end{cases} \qquad \text{Eq.(4.37)}$$

Les polynômes $F_1'(z)$ et $F_2'(z)$ étant symétrique et antisymétrique, cette équation est directement issue de la relation $A(z) = (F_1'(z) + F_2'(z))/2$.

4.3 Estimation de la période du fondamental et décision de voisement

4.3.1 Introduction

La période $P=1/f_0$ du fondamental (appelé communément le pitch) joue un rôle prépondérant pour la transformation du signal de parole particulièrement dans la phase de resynthèse ; l'oreille est en effet très sensible à ses variations, lesquelles constituent la prosodie (ce qu'est la mélodie à la musique)

L'estimation du pitch est bien sur liée à la localisation des tranches voisées du signal de parole. En général, lorsqu'une méthode ne fournit pas de valeur admissible pour P, on décrète que le signal est non voisé ; d'autres critères peuvent cependant êtres utilisés.

La période P est comprise généralement entre 2 et 17 ms, soit une fréquence f_0 située entre 80 et 250 Hz pour les voix masculines et entre 150 et 500Hz pour les voix féminines.

L'estimation du pitch est en pratique une tâche difficile pour de nombreuses raisons telle que la non-stationnarité du signal, certaines irrégularités dans l'excitation glottique ou encore une interaction avec le premier formant.

De très nombreuses méthodes pour l'estimation du pitch ont été proposées ; elles comportent généralement trois étapes :

- Un prétraitement du signal.
- L'estimation proprement dite.
- Un post-traitement.

Le pré-traitement réalise en général une compression des données, par exemple une décimation dans le rapport 5 à 1, pour rendre possible une estimation en temps réel.

L'estimation proprement dite peut être effectuée dans le domaine temporel ou sur le résultat d'une analyse court-terme (fonction d'autocorrèlation ou densité spectrale) effectuée par le pré-traitement.

Le post-traitement est un filtrage dont le rôle consiste à corriger les erreurs éventuelles et à lisser le contour du pitch en tenant compte de valeurs à priori.

Quelques méthodes de complexités diverses vont être brièvement décrites ; beaucoup de travaux sont menés à ce jour pour améliorer les techniques et méthodes de détection de pitch du fait de son importance en traitement de la parole. [Son68], [Ros74],[sun02]…

Les algorithmes de détection de pitch se subdivisent en méthodes qui opèrent dans le domaine de temporel, dans le domaine fréquentiel ou les deux.
Certaines méthodes de détection de pitch utilisent la détection et la synchronisation de singularités dans le domaine temporel à l'exemple des IFG (Instants de Fermeture de Glotte) dont le cycle de répétition détermine la période du fondamental.
D'autres méthodes dans le domaine temporel emploient des fonctions d'autocorrélation pour détecter des similitudes entre formes d'ondes.
Une autre famille de méthodes fonctionne dans le domaine fréquentiel et utilise une modélisation spectrale du signal d'entrée permettant sa description en tant que superposition de sinusoïdes, dont le rapport des fréquences détermine la fréquence fondamentale du signal.

Nous décrivons brièvement dans ce qui suit quelques méthodes « basiques » et nous terminerons par d'autres méthodes plus élaborées reflétant les recherches les plus récentes dans le domaine.

4.3.2 Taux de passage par zéro

Technique simple consistant à compter le nombre de fois que le signal croise le niveau de référence 0 . Cette technique est très simple et peu coûteuse mais n'est pas très précise. En fait en traitant les signaux fortement bruités ou les signaux harmoniques où les secondaires sont plus forts que le fondamental, la méthode n'est pas très performante

4.3.3 Méthode basée sur la fonction d'autocorrèlation.

On calcule la fonction d'autocorrèlation sur une tranche de N échantillons qui recouvre plusieurs périodes du fondamental

$$r(k)=\sum_{n=1}^{N-1-k} s(n)s(n+k) \quad , \text{k} = 0 ,\ldots, \text{K} \qquad \text{Eq. (4.27)}$$

A titre d'exemple, pour f_e = 10 Khz on peut choisir N= 300 et K=150 et on recherche un maximum prononcé dans l'intervalle (0,K).

La fonction d'autocorrélation contenant une information surabondante sur le signal, le coût du calcul est alors très important, on est donc amenés pour réduire ce coût à ne retenir de chaque échantillon s(n) que la partie qui excède un certain seuil (center clipped signal) ; le seuil est défini par rapport à la plus grande valeur en module de s(n) dans la tranche analysée (par exemple 60%). Voir figure 4.8

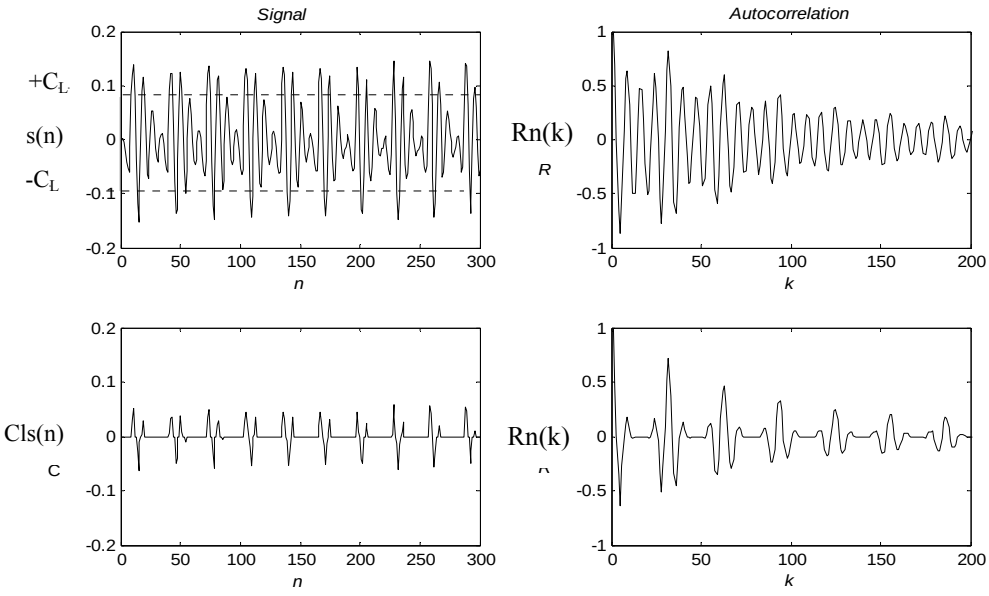

Fig. 4.8 Illustration de la méthode d'autocorrelation

En haut signal original et sa fonction d'autocorrelation.
En bas signal dont on a supprimé le noyau $|s| < C_L$ (zone grisée sur la figure) et sa fonction d'autocorrelation

4.3.4 Méthode AMDF

La méthode connue sous le sigle AMDF (Average Magnitude Difference Function) pour la mesure de la périodicité d'une tranche de signal de parole voisée est définie par la relation :

$$D(k)=\frac{1}{N}\sum_{n=1}^{N}\left|s(n)-s(n-k)\right| \qquad \text{, k = 0, 1, …, K} \qquad \text{Eq. (4.28)}$$

où s(n) sont les échantillons de parole et s(n-k) les échantillons décalés de k secondes.

Si le signal analysé est parfaitement périodique, on devrait obtenir D(i) = 0 pour i = 0,1,… la pratique montre que l'estimation de f_0 par la recherche d'un creux dans la suite D(k) reste assez aisée malgré la non stationnarité.

Il existe par ailleurs une relation approchée entre D(k) et r(k) obtenue d'après[Ros74] comme suit :

On part de l'inégalité

$$\frac{1}{N}\sum_{n=0}^{N-1}\left|u(n)\right| \;\le\; \left[\frac{1}{N}\sum_{n=0}^{N-1}u^{2}(n)\right]^{1/2} \qquad \text{Eq. (4.29)}$$

basée sur l'inégalité de Schwarz.

L'équation 4.28 peut alors être approchée par :

$$D(k)=\frac{1}{N}\sum_{n=1}^{N}\left|s(n)-s(n-k)\right|\cong\beta_k\left(\frac{1}{N}\sum_{n=1}^{N}\left(s(n)-s(n-k)\right)^2\right)^{1/2} \qquad \text{Eq. (4.30)}$$

où le coefficient β_k est un facteur scalaire dont la valeur peut être défini de manière analytique ou expérimentale selon le type de distribution des séquences de parole analysées.

En développant le dernier terme de l'équation précédente on obtient :

$$D(k)\;\cong\;\beta_k\left(\frac{1}{N}\sum_{n=1}^{N}s^2(n)+\frac{1}{N}\sum_{n=1}^{N}s^2(n-k)-\frac{2}{N}\sum_{n=1}^{N}s(n)s(n-k)\right)^{1/2} \qquad \text{Eq. (4.31)}$$

et en définissant la fonction d'autocorrélation par :

$$r(k)=\frac{1}{N}\sum_{n=1}^{N}s(n)s(n-k) \qquad \text{Eq. (4.32)}$$

il ressort de l'équation 4.31 que le troisième terme représente –2r(k).

Par ailleurs pour un processus stationnaire les deux premiers termes ne sont que la fonction d'autocorrelation évaluée à k=0 soit :

$$r(0)=\frac{1}{N}\sum_{n=1}^{N}s^2(n)\cong\frac{1}{N}\sum_{n=1}^{N}s^2(n-k) \qquad \text{Eq. (4.33)}$$

l'équation 4.31 sous condition de stationnarité s'écrira alors :

$$D(k) \cong \beta_k\left[2\left(r(0)-r(k)\right)\right]^{1/2}$$

4.3.5 La méthode du filtre inverse SIFT

La méthode connue sous le sigle SIFT (Simplified Inverse Filter Tracking algorithm) est basée sur l'examen de la fonction d'autocorrèlation du résidu LPC

Afin de faciliter le calcul temps réel, on opère une décimation dans un rapport 5 à 1 après passage dans un filtre passe-bas voir fig 4.9 pour illustration.

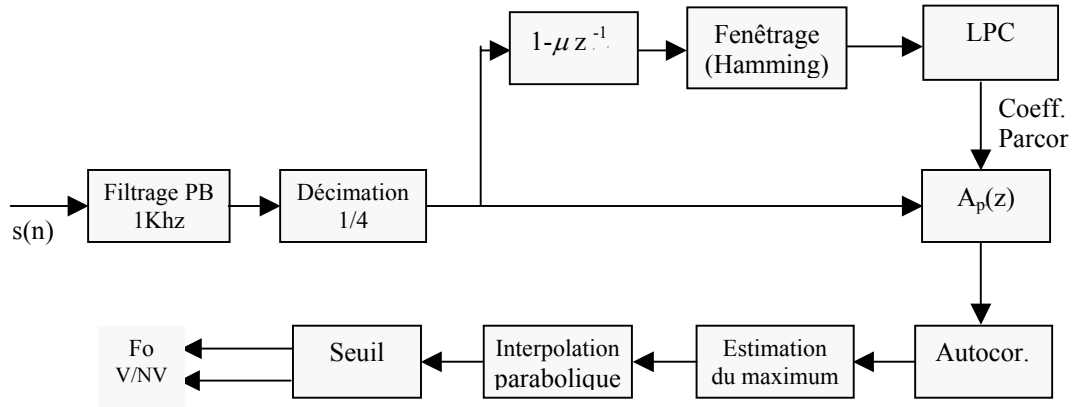

Figure 4.9. Schéma bloc Méthode du SIFT

Après une préaccentuation et une pondération, on procède à une analyse LPC (p=4) pour définir le filtre inverse. La méthode de l'autocorrèlation, ou éventuellement celle de l'AMDF, est appliquée au résidu e(n). La position du maximum doit être affinée par interpolation parabolique car la période d'échantillonnage en ce point de la chaîne doit être de 0.5 ms.

Pour être accepté, un maximum (rapporté à r(0)) doit dépasser un seuil variable avec k dont le rôle consiste à éliminer un maximum non lié à la période du pitch ; la loi représentée à la figure 4.10 a été proposée en [MaG76].

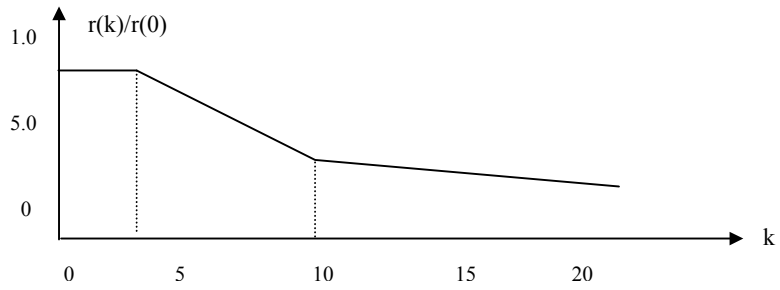

Figure 4.10 Seuil d'acceptabilité pour la méthode du SIFT

4.3.6 Méthode du cepstre

L'estimation de la période du pitch peut être effectuée sur le cepstre réel, celui-ci est calculé comme illustré à la figure 4.11

Le signal est découpé en tranches de 300 échantillons avec recouvrement de 10 échantillons ; chaque bloc est pondéré par une fenêtre temporelle avec n_0=20.

Les maxima sont recherchés dans un intervalle de 2 à 15 ms ; un maximum est accepté lorsqu'il excède la courbe du seuil représentée à la figure 4.12, ce qui permet en principe d'éviter le phénomène de doublement du pitch (estimation erronée de 2fo au lieu de fo).

Si aucun maximum n'excède le seuil, la tranche est décrétée non voisée. Cette décision est confirmée par un filtre logique si les deux tranches précédentes sont non voisées ; dans le cas contraire, le seuil est baissé de 25%.

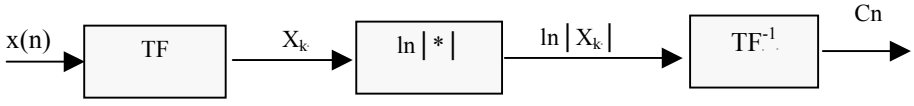

Figure 4.11 Calcul du cepstre réel

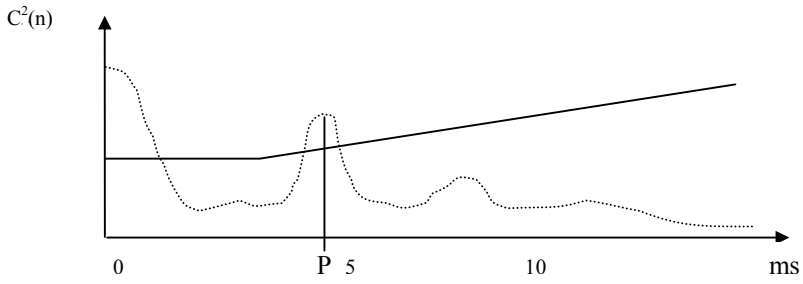

Figure 4.12 - Seuil d'acceptabilité pour la méthode du cepstre

4.3.7 Algorithme « SHRP »

A pitch determination algorithm based on Subharmonic-to-Harmonic Ratio (SHR)

Cet algorithme relativement récent (2002) voir référence [Sun02] est souvent cité dans la littérature est basé comme son nom l'indique sur le rapport sous-harmoniques à harmoniques et est basé sur des considérations perceptives du signal de parole. En effet les résultats de tests d'écoute menés par l'auteur ont montré que la perception du pitch est fortement liée au SHR c'est à dire l'amplitude du rapport sous-harmoniques à harmoniques. Sa procédure est mise à disposition par l'auteur *. Nous l'avons utilisé pour effectuer les calculs statistiques sur la base de données TIMIT.(voir tableaux 5.13 et 5.14).

---

* **Xuejing Sun**

http://www.mathworks.com/matlabcentral/fileexchange/loadFile.do?objectId=1230&objectType=file

4.4. Marquage de pitch

4.4.1 Introduction.

Les méthodes reposant sur le principe de synchronisation avec le fondamental sont utilisées pour réaliser des modifications temporelles ou fréquentielles d'un signal de parole ou pour mettre en œuvre des systèmes de synthèse à partir du texte.
Ces méthodes nécessitent au préalable un marquage des périodes du fondamental. TD-PSOLA est basée sur la décomposition du signal de parole en fenêtres recouvrantes synchronisées sur les périodes du fondamental. Les marques de synchronisation indiquent les centres de ces fenêtres. Les modifications consistent à manipuler les marques d'analyse pour générer de nouvelles marques de synthèse. Cela correspond à la duplication ou l'élimination de fenêtres dont l'écartement peut être modifié.
La principale exigence, commune aux techniques de décomposition du signal en courtes fenêtres, est de garder la cohérence mutuelle des emplacements des marques pour préserver la structure temporelle originale du signal étudié. Par conséquent, il est crucial d'obtenir un marquage précis des périodes du fondamental car il influe directement sur la qualité du signal.

De nombreuses méthodes de marquage ont été décrites dans la littérature [LaC98].
Elles sont généralement basées sur la recherche d'évènements précis dans le signal de parole :
Instants de fermeture glottale IFG (ou en anglais IGC : Instant Glottal Closure), extrema du signal, instants d'excitation des modèles LPC, etc.…Ces techniques peuvent souffrir d'une certaine « rigidité », elles peuvent forcer le marquage d'échantillons qui satisfont à un critère numérique mais à s'éloigner des marques voisines par rapport à la période du fondamental.

L'algorithme de marquage automatique de pitch sur un signal de parole continue, proposé par *Vladimir Goncharoff and Patrick Gries* [GoG98] a particulièrement attiré notre attention par son efficacité nous l'avons d'ailleurs utilisé dans le cadre de nos expérimentations.

Une particularité pour cette méthode est que les marques de pitch sont régulièrement espacées, même pendant les silences ou sur les régions non voisées du signal où il n'y a aucune composante périodique. Ainsi des décisions de voisement n'ont pas besoin d'être prises et des marques de pitch sont assignées à travers le signal entier. Cette méthode a des applications dans l'analyse de la parole par PSOLA, où non seulement la période de pitch mais également les centres d'impulsion de pitch doivent être exactement marqués.

PSOLA et d'autres algorithmes pitch-synchrones de traitement des signaux de parole exigent une connaissance de la fréquence de pitch et de la phase *de pitch* c.-à-d., les endroits des marques placées sur les maxima de la courbe d'énergie à court terme de chaque impulsion de pitch. Pour cette méthode les auteurs ont choisi de définir les endroits désirés des marques de pitch comme suit:

- Sur des segments (voisés) quasi-périodiques de la parole, une marque de pitch doit être placée sur l'un des extremas d'amplitude de chaque cycle de pitch de manière qu'avec les marques voisines de pitch, il résulte une séquence de marques lissée.
- Pendant les segments de parole n'ayant aucune composante périodique dans sa forme d'onde, les marques de pitch doivent être placées de manière que la séquence de marques obtenues soit lissée avec les marques issues des segments voisés adjacents.

Cette méthode ne fait aucune distinction entre les régions voisées et non voisées; des marques de pitch sont continûment placées sur le signal de parole entier. Le schéma 1 montre une section de la parole qui commence comme voisé, devient non voisé, puis finit comme voisée, avec les marques de pitch trouvées par cet algorithme.

L'algorithme de marquage de pitch est constitué de deux blocs indépendants: évaluation de période de pitch, et évaluation de phase de pitch.

Dans le premier bloc la période de pitch est grossièrement estimée, sans souci des positions de marque de pitch sur les segments voisés, puis interpolée en tant que fonction continue de segments non voisés.. Dans le deuxième bloc cette évaluation brute de période de pitch est raffinée et des marques de pitch sont placées aux pics d'amplitude de forme d'onde pour compléter le procédé. Chacun de ces deux blocs est basé sur une routine de programmation dynamique.

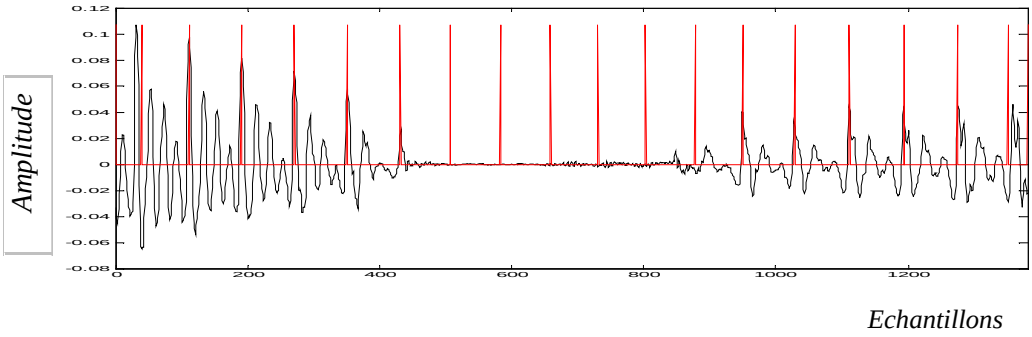

Figure 4.13. *Signal de parole et marques placées par l'algorithme.*
(sur les régions voisées et non voisées)

4.4.2 Estimation du chemin optimal par programmation dynamique

Une routine de programmation dynamique est utilisée pour calculer un chemin optimal de la première à la dernière colonne d'une matrice rectangulaire, sous des contraintes de pente maximum. Etant donné une matrice A de dimensions N par M il est défini un vecteur p qui indique un index de ligne pour chaque colonne**.** La somme de valeurs de coefficients le long du chemin à travers la matrice est donnée par :

$$E=\sum_{m=1}^{M} A(p(m),m) \qquad \text{Eq 4.31}$$

Cette routine de programmation dynamique détermine le vecteur optimal **p** pour lequel *E* est maximum, sous la contrainte :

$$\left|p(m)-p(m-1)\right| \le S_{\max} \qquad \text{pour } m \quad 2 \le m \le M \qquad \text{Eq 4.32.}$$

et où S_{max} représente est la pente maximale permise pour le chemin.
Cette version de l'algorithme a été efficacement implémentée sous "MATLAB", le code est mis à disposition par les auteurs *Vladimir Goncharoff and Patrick Gries* [GoG98] par E-mail. (*)
Une version modifiée par Oytun Turk - 01.11.2002 supporte n'importe quel fréquence d'échantillonnage.

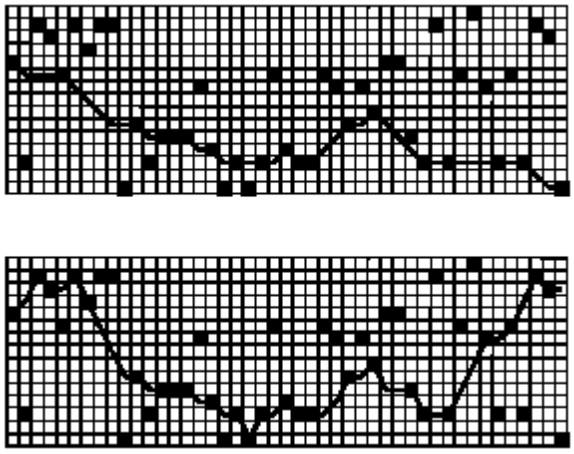

Fig 4.14. - Illustration de l'algorithme de programmation dynamique.[GoG98]

Le chemin sur la matrice du haut est calculé avec Smax = 1 alors que pour la matrice du bas Smax=2. Le chemin optimal est calculé en passant par le maximum de carrés sombres.

La programmation dynamique dans le passé a été appliquée aux diverses applications de traitement de signal, y compris le marquage de pitch.
Cet algorithme à deux phases utilise la programmation dynamique pour marquer de façon très efficace les impulsions de pitch sans avoir à détecter de zones de voisement / non-voisement. En outre, des marques de pitch sont placées de façon continue même pendant le silence ou sur les segments non voisés. L'algorithme peut facilement être appliqué au marquage de pitch en temps réel pour les signaux de parole dans les systèmes de conversion de voix.

* Vladimir Goncharoff, Department of Electrical and Computer Engineering College of EngineeringUniversity of Illinois at Chicago mail : goncharo@eecs.uic.edu

4.5 Erreurs et indices de performance

4.5.1 Mesures objectives

Il y a deux sortes d'erreurs particulièrement importantes : l'erreur de transformation *Ect* entre les vecteurs cible et transformé, et l'erreur inter-locuteurs *Esc* entre les vecteurs source et cible. Ces erreurs ne peuvent pas être mesurées directement, mais peuvent être approchées empiriquement par une distance moyenne entre les vecteurs spectraux.

Etant donnée une distance d et trois suites de vecteurs spectraux $S = [S_1, S_2,\ldots, S_N]$, $C = [C_1, C_2,\ldots,C_N]$ et $T = [T_1,T_2,\ldots, T_N]$, correspondant respectivement aux vecteurs source, cible et transformés, les erreurs décrites précédemment s'écrivent alors :

$$E_{ct} = \frac{1}{N}\sum_{i=1}^{N} d(C_i,T_i) \qquad \text{Eq.4.33}$$

$$E_{sc} = \frac{1}{N}\sum_{i=1}^{N} d(S_i,C_i) \qquad \text{Eq. 4.34}$$

Pour pouvoir mesurer l'apport de la transformation, on utilise une erreur normalisée (*EN*), qui est le rapport des deux erreurs décrites précédemment, ce rapport représentant une mesure normalisée de la proximité entre les paramètres convertis et ceux de la cible :

$$E_N = \frac{E_{CT}}{E_{SC}} \qquad \text{Eq.4.35}$$

La distance la plus connue et la plus utilisée est la distance euclidienne :

$$d(A,B) = \sqrt{\sum_{i=1}^{p} |a_i - b_i|^2} \qquad \text{Eq.4.36}$$

Dans l'expression de la distance euclidienne nous supposons implicitement que, perceptuellement, les différentes composantes des vecteurs contribuent, de

manière identique à la distance globale. En pratique, certaines composantes ont une plus grande importance. Par exemple, dans le cas des coefficients LSFs, l'oreille humaine est beaucoup plus sensible à une variation des premiers coefficients qu'à la même variation des derniers.

Pour répondre à ce problème, il est nécessaire de faire intervenir une pondération pour bien refléter ce phénomène. L'expression générale d'une distance quadratique pondérée est la suivante :

$$d(A,B) = (A-B)^T W(A,B) \qquad \text{Eq.4.37}$$

où *W* est une matrice carrée définie positive. Lorsque *W* est l'inverse de la matrice de covariance des vecteurs spectraux, on parle alors de la distance de Mahalanobis.

Cependant, une distance quadratique, pondérée ou non, entre deux vecteurs de cepstre discret, ou deux vecteurs LSF n'a pas la même signification dans le domaine spectral.

En d'autres termes, le fait que la distance entre deux vecteurs de coefficients cepstraux soit plus petite que la distance entre deux vecteurs de coefficients LSF ne signifie pas que les deux spectres synthétisés par les deux vecteurs cepstraux seront plus proches perceptuellement que ceux synthétisés par les deux vecteurs LSFs. Pour respecter la sensibilité de l'oreille aux distorsions spectrales, des distances spectrales particulières ont été utilisée. Ces distances sont fonction des densités spectrales des signaux originaux et transformés, ce qui fait d'eux une jauge mieux adaptée pour une comparaison de la qualité de conversion du cepstre et des LSFs.

Gillet dans [Gil03] et dans le cadre de l'évaluation de la transformation de l'enveloppe spectrale a estimé l'erreur entre deux suites de vecteurs LSF par :

$$E_{LSF}(A,B)=\frac{1}{M}\sum_{m=1}^{M}\sqrt{\frac{1}{P}\sum_{i=1}^{P}(L_A^{m,i}-L_B^{m,i})^2} \qquad \text{Eq. 4.38}$$

où M est le nombre de trames , P l'ordre LPC et $L^{m,i}$ est le $i_{ème}$ composant du vecteur LSF de la trame m.

Cette définition de l'erreur permet alors d'évaluer la différence entre deux locuteurs à travers l'erreur inter-locuteurs $E_{LSF}(S(n),C(n))$, où C(n) représente les LSF du locuteur cible.

Si les LSF prédits pour le locuteur cible sont notés $\hat{C}(n)$ l'erreur de transformation sera définie par la différence entre les LSF actuels et prédits

L'indice de performance des LSF suggéré par Kain[01] pour évaluer la qualité de la transformation s'écrit alors :

$$P_{LSF} = 1 - \frac{E_{LSF}(C(n),\hat{C}(n))}{E_{LSF}(C(n),S(n))} \qquad \text{Eq. 4.39}$$

une valeur de $P_{LSF} = 1$ indiquera alors que la sortie du système de transformation est identique à la cible (cas idéal).

Ce sera cet indice qui sera utilisé pour évaluer la performance de notre système quant à la transformation de l'enveloppe spectrale.

Pour l'évaluation objective de la conversion de pitch il est suggéré dans [Enn05] d'utiliser une distance quadratique moyenne entre les valeurs de pitch normalisé appelée DPN (pour Distorsion de Pitch Normalisé) définie comme suit :

Si on définit la valeur de pitch normalisée par

$$F_{\log} = \log\left(\frac{F_0}{\hat{F}_0}\right) \qquad \text{Eq. 4.40}$$

où F_0 est la fréquence fondamentale en Hz et $\hat{F}$ est la moyenne des valeurs de pitch sur toute la base d'apprentissage.

Alors la DPN s'écrira :

$$DPN = \sqrt{\frac{1}{N}\sum_{n=1}^{N}\left(\frac{g_n^y - \hat{g}_n^y}{g_n^y - g_n^x}\right)^2} \qquad \text{Eq.4.41}$$

où g_n^x, g_n^y, *et* $\hat{g}_n^y$ sont les valeurs de la fréquence fondamentale source, cible et convertie respectivement et normalisées conformément à l'équation 4.40.

4.5.2 Mesures subjectives.

Pour évaluer un système de conversion de voix il communément utilisé des tests d'écoute. Le test d'écoute le plus utilisé pour valider un système de conversion de voix est le test ABX , c'est un test au cours duquel des auditeurs jugent si le son prononcé par un locuteur « X » est plus proche du locuteur « A » ou du locuteur « B ». Dans le cas de la conversion inter-genre, c'est à dire, la conversion homme/femme et femme/homme ce type de test se révèle inadéquat du fait que le système arrive à simuler le genre cible, la voix convertie est jugée plus proche de la cible que de la source.

Pour juger de la qualité des phrases prononcées il peut alors faire appel au test MOS (Mean Opinion Score) à cinq niveaux. Cette échelle va de un, pour la plus mauvaise qualité, jusqu'à cinq pour une qualité excellente comme indiqué sur le tableau suivant :

Score	Qualité de la parole	Niveau de distorsion
5	Excellente	Imperceptible
4	Bonne	Perceptible mais non gênant
3	Moyenne	Perceptible et faiblement gênant
2	Médiocre	Gênant mais supportable
1	Mauvaise	Gênant et insupportable

Tableau 4.15 Description de scores dans les tests MOS.

Chapitre 5

Mise en oeuvre de la conversion de voix

5. Mise en oeuvre de la conversion de voix

5.1 - Introduction

Dans le cadre de notre travail, l'objectif de la transformation du signal de parole est principalement la conversion de voix d'un locuteur source vers un locuteur cible.

La méthode PSOLA utilisée pour la transformation temporelle ou fréquentielle du signal de parole c'est à dire la modification de la durée ou de la hauteur ne se prête pas telle quelle à la conversion de voix c'est ce qui nous a amenés à rechercher une plate forme appropriée à son utilisation en conversion de voix où la transformation du signal de parole consiste en la modification des trois paramètres prosodiques suivants: l'énergie, la durée et la fréquence fondamentale mais aussi la modification de l'enveloppe spectrale permettant d'approcher aussi le timbre du locuteur cible.
La plate forme devrait être basée sur une modélisation capable de nous fournir les possibilités de conversion adéquates à travers la transformation ou la transplantation des paramètres caractéristiques indépendamment les uns des autres.

Par ailleurs la méthode PSOLA basée sur la technique de la TFCT (Transformée de Fourier à Court Terme) qui se limite du point de vue de l'analyse spectrale du signal de parole à fournir les caractéristiques de la source et du filtre en une seule information mixée est inadéquate à son exploitation en conversion de voix.

L'analyse de la parole par la technique LP est devenue la technique prédominante pour l'estimation des paramètres de base du signal de parole que se soit en analyse- synthèse, en codage ou autres domaines du traitement de la parole.

La plate forme LP adoptée nous permet donc de procéder à la conversion des différents paramètres de même que la méthode de transformation PSOLA trouvera une application selon le schéma LP-PSOLA cité dans le chapitre 2.

5.1. Transformation du signal par la méthode PSOLA.

Nous présentons dans cette première partie quelques résultats préliminaires de transformations du signal de parole par la méthode PSOLA.

On rappelle que les signaux de parole utilisés sont issus de la base de données TIMIT (voir Annexe A) et sont échantillonnés à 16 KHz.

5.1.1 Etapes de la transformation PSOLA.

Les étapes de transformation sont décrites par le schéma bloc de la figure suivante :

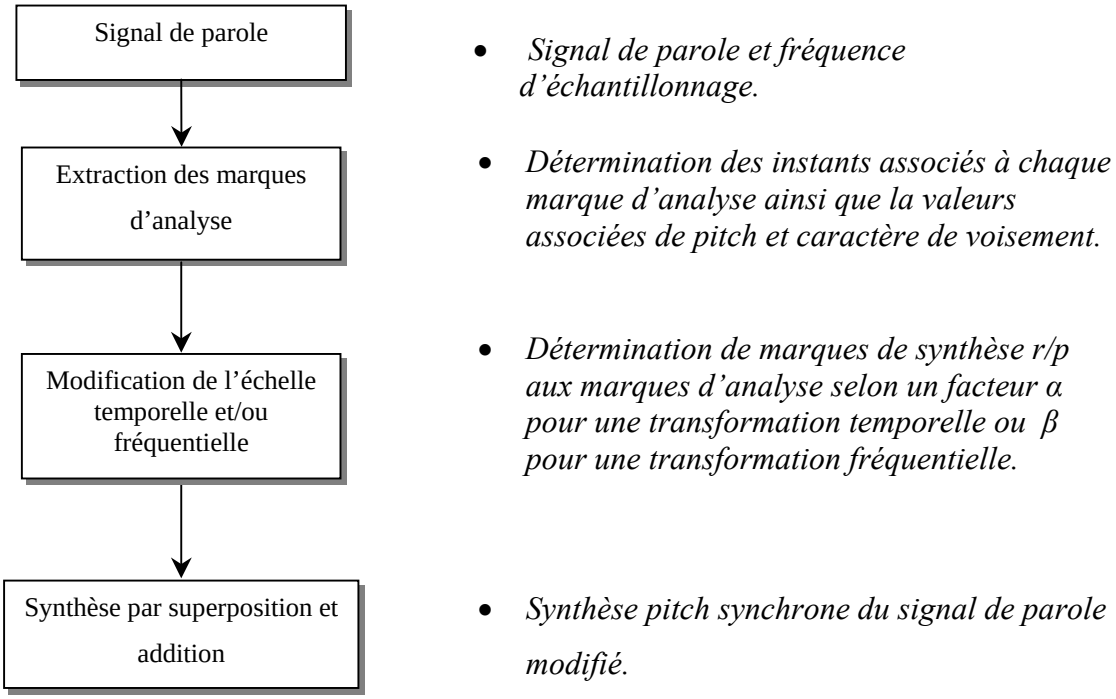

Figure 5.1 – Schéma bloc de la transformation PSOLA

5.1.2 Modification de l'échelle temporelle (Expansion/compression) :

L'exemple présenté a pour signal original la phrase équilibrée :
Train\Dr1\Medro\sx24 émise par un locuteur masculin (Although always alone, we survive.)

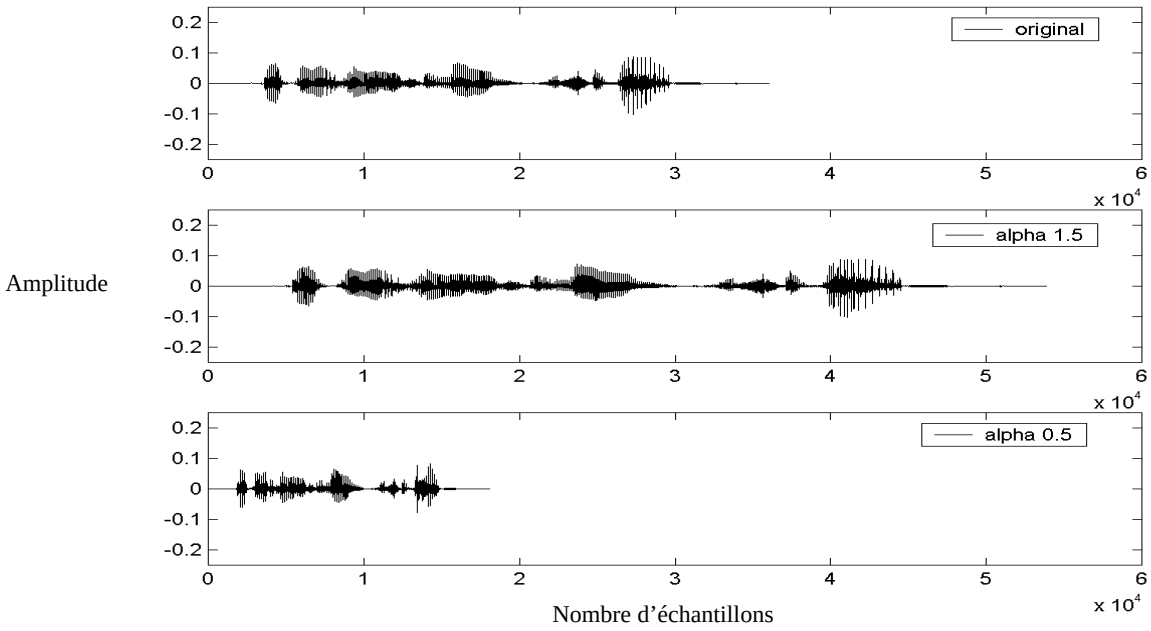

Figure 5.2 Expansion, compression par TD-PSOLA

Cet exemple montre une transformation temporelle par un facteur alpha=1.5 (ralentissement) puis un facteur alpha = 0.5 (compression).
On notera ici que le ralentissement ou la compression sont effectués en conservant la valeur du pitch.
D'après les tests d'écoute effectués sur les multiples phrases de la base de données, un facteur alpha inférieur à 0.5 rend la phrase inintelligible du fait de la trop grande vitesse d'élocution.
Par ailleurs il existe des limites aux rapports de compression importants qui peuvent dénaturer le signal et cela s'explique par exemple pour un phonème voisé contenant quatre périodes de pitch, une compression au-delà de 25%, réduit ce phonème à moins d'une période de pitch, éliminant le caractère de voisement de ce phonème.
Et au delà d'un facteur 2, une faible réverbération est perceptible alors que les signaux transformés sont de très bonne qualité. Par ailleurs le pitch est bien conservé.

5.1.3 Modification de l'échelle fréquentielle :

C'est l'opération duale de la précédente où on modifie la valeur de la fréquence fondamentale tout en conservant la durée du signal. Le même signal test est utilisé pour cette transformation de l'échelle fréquentielle. On notera par une simple analyse du signal transformé que la valeur du pitch a été modifiée alors que la durée est conservée.

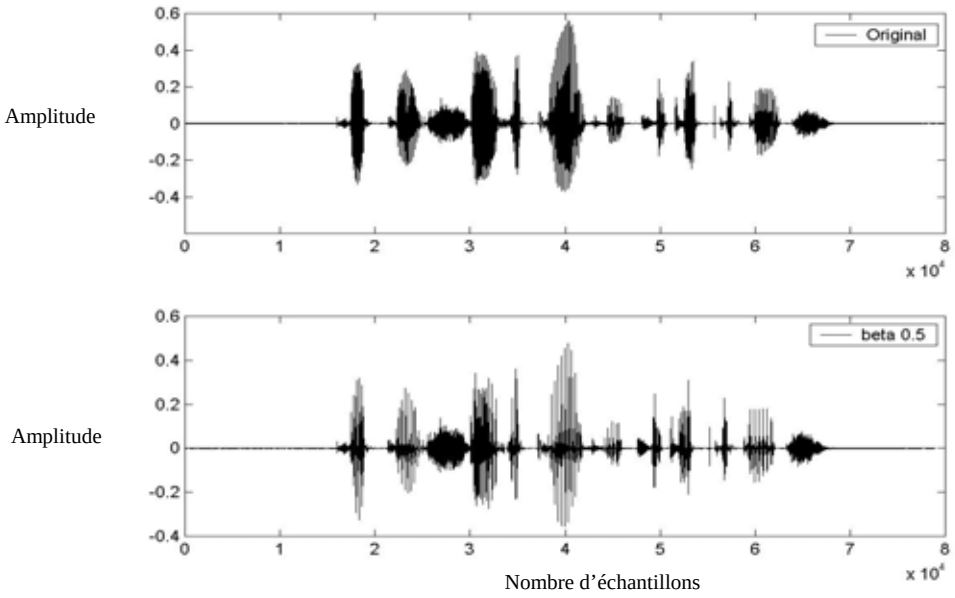

Figure 5.2 Modification du pitch par TD-PSOLA

Conclusion

La méthode PSOLA (TD-PSOLA) s'est avérée être un moyen simple et efficace pour la modification du pitch et de la durée des signaux de parole mais du fait que la méthode procède dans le domaine temporel, l'enveloppe spectrale ne peut être contrôlée de manière efficace.

La méthode conviendrait parfaitement à une utilisation en synthèse par concaténation où des segments de parole préenregistrés sont soumis à une transformation prosodique (pitch et/ou durée) ciblée.

5.2. Mise en œuvre de la conversion de voix.

5.2.1 Introduction.

Le programme élaboré dont le schéma synoptique est représenté par la figure 5.4 a pour objectif une conversion de voix basée sur une transformation conjointe de l'enveloppe spectrale et de la fréquence fondamentale.
L'apprentissage de notre programme se limite à une paire de phrases. La 1ére phrase est émise par un premier locuteur appelé locuteur « source » et un deuxième locuteur appelé locuteur « cible ». Le locuteur source émet une phrase « test » que le programme se chargera de convertir selon les règles de transformation apprises de manière qu'à l'écoute elle semble avoir été émise par le locuteur cible.
Les grandes lignes de cette conversion sont principalement les étapes de « mapping » (ou mise en correspondance) et l'extraction des paramètres de transformation.

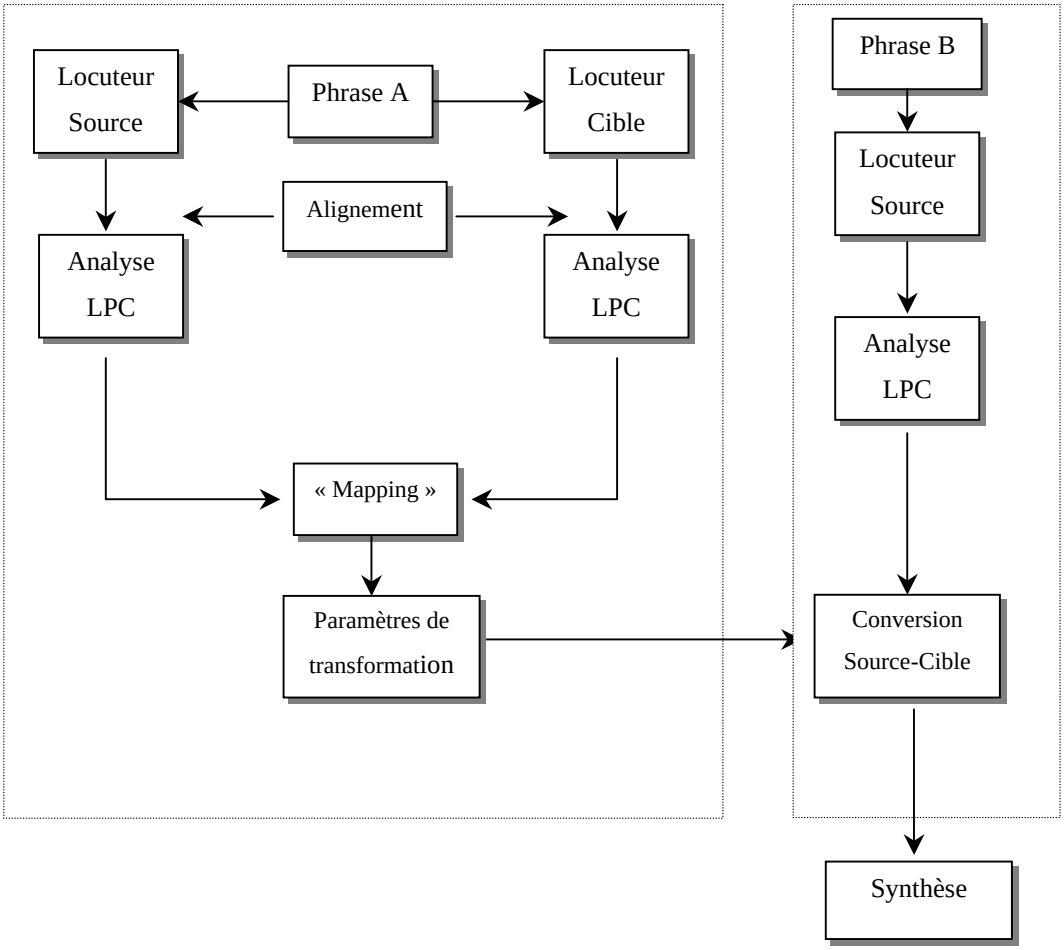

Figure 5.4- Schéma global pour la conversion

La figure 5.5 schématise le modèle de production de la parole qui nous servira de référence quant au module de synthèse du schéma synoptique de la conversion ainsi que la manière d'effectuer la conversion.

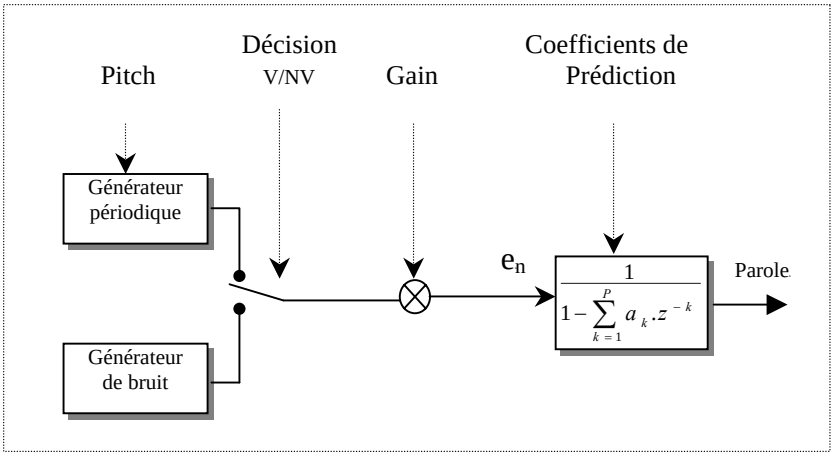

Figure 5.5. : Modèle paramétrique de production de la parole

5.2.2 Transformation de l'enveloppe spectrale.

5.2.2.1 Extraction des paramètres.

Après l'analyse LPC utilisant la méthode de l'autocorrelation sur les trames du signal de parole on dispose pour chaque locuteur source et cible des coefficients de prédiction notés dans la suite respectivement par a_srce et a_cble, des valeurs de gain ainsi que du signal résiduel. Les coefficients de prédictions sont ensuite convertis en coefficients LSP notés a_srce_LSP et a_cble_LSP

A ce niveau de notre expérimentation nous devions trancher quant au choix de :

- La méthode d'alignement.
- L'ordre de l'analyse LPC.
- La régression sur les coefficients de prédiction.

Ceci nous a amenés à effectuer des mesures de performance comme suggéré dans le chapitre 4.

5.2.2.2 Alignement DTW ou PSOLA ?

L'étape d'alignement est un passage obligé comme pré-traitement de la phrase d'apprentissage émise par les deux locuteurs du fait des différences de vitesse d'élocution. Celle ci est généralement effectuée par la méthode DTW. Nous avons voulu à ce niveau tirer profit des caractéristiques de la technique PSOLA en remplacement de la DTW. Ceci à été motivé par le fait que les phrases alignées par DTW présentaient à l'écoute une certaine réverbération qui pourrait fausser au départ la qualité de la conversion, surtout pour les phrases présentant une différence relativement importante de durée.

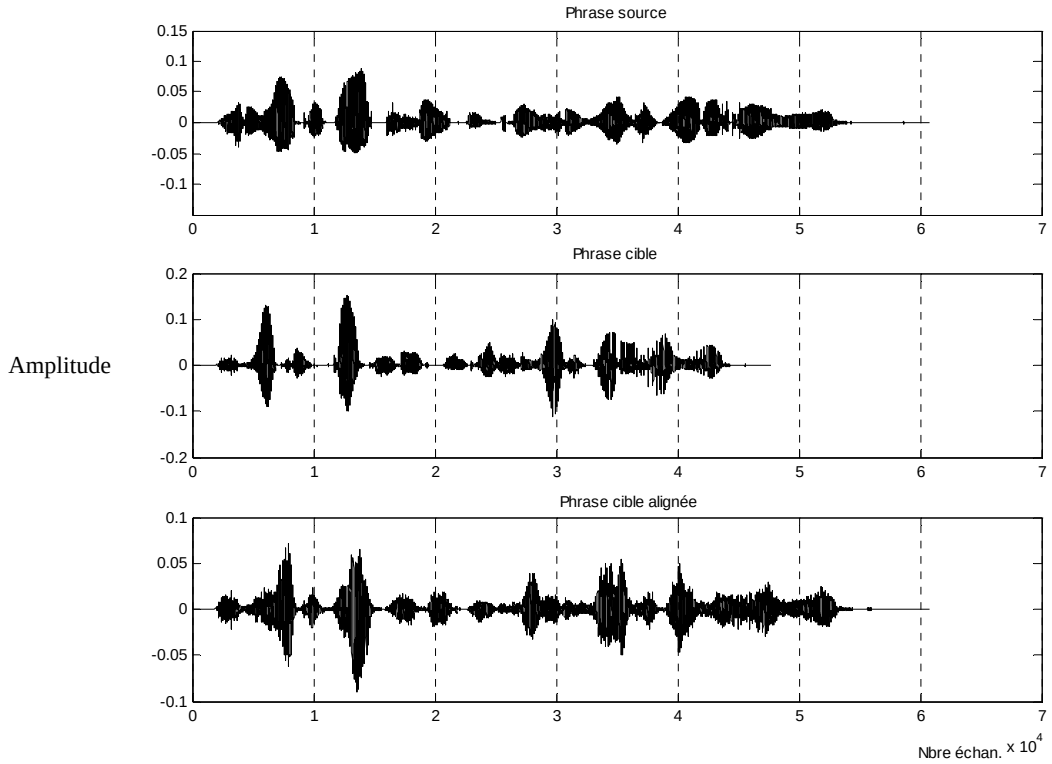

Figure 5.6 - Alignement de phrases par DTW

Malheureusement l'application de la méthode PSOLA ne pouvait aboutir comme cela était attendu qu'à une simple « mise à l'échelle » ceci est illustré par les figures 5.6 et 5.7 (sur la figure 5.7 b on a supprimé les silences précédant et terminant la phrase) bien que la qualité de la phrase restituée par la méthode PSOLA soit de loin de meilleure qualité que celle obtenue par DTW.

Retenant le principe de suppression du silence de début et de fin de phrase nous avons opté dans notre travail pour un alignement des paires de phrases utilisées par l'algorithme DTW

sur la base des coefficients LSP. Les coefficients LSP ne sont pas les seuls candidats à la représentation de l'enveloppe spectrale, les coefficients MFCC utilisés dans par Gillet [Gil03] et Ennejary [Enn05] dans le cadre de transformation de pitch en conversion de voix mais du fait que les coefficients LSP supportent mieux l'interpolation [Pal95], comme cela à déjà été cité, nous avons donc préféré utiliser ceux-là au lieu des coefficients cepstraux..

Notons par ailleurs que dans [Ina03] toujours dans le cadre de transformation de pitch en conversion de voix l'auteur a utilisé un alignement basé sur une transcription phonétique des phrases source et cible, cet alignement est adopté dans les travaux les plus récents et donne d'après les auteurs des résultats très performants.

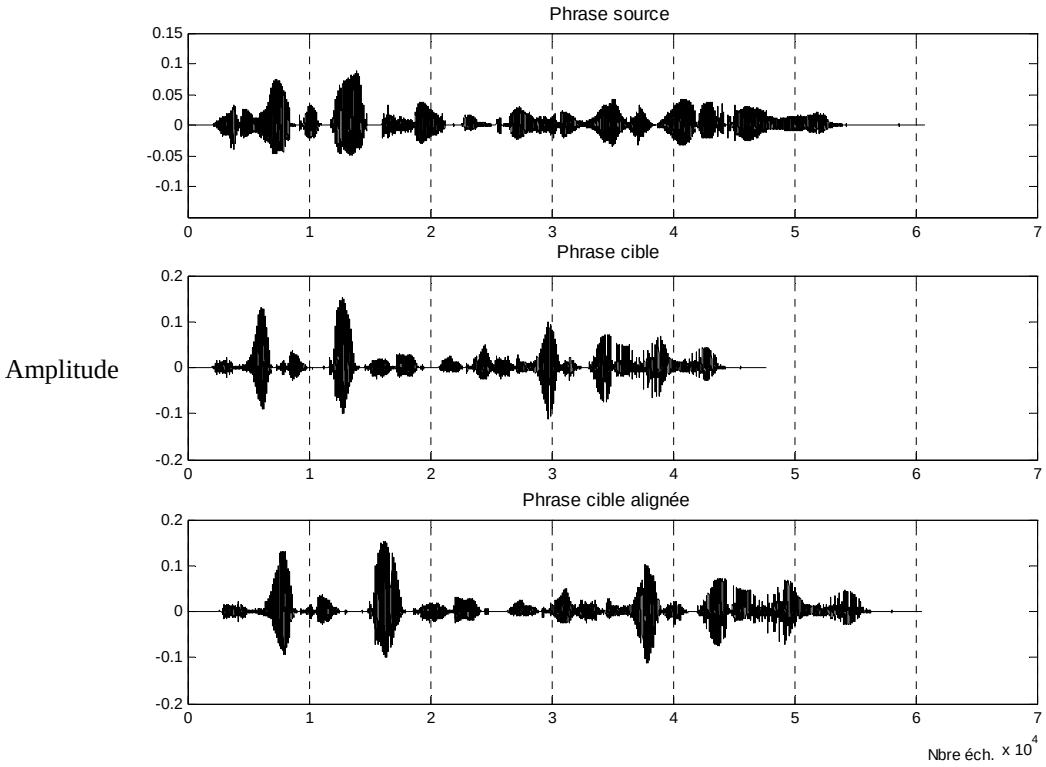

Figure 5.7a - « Alignement » de phrases par la méthode PSOLA

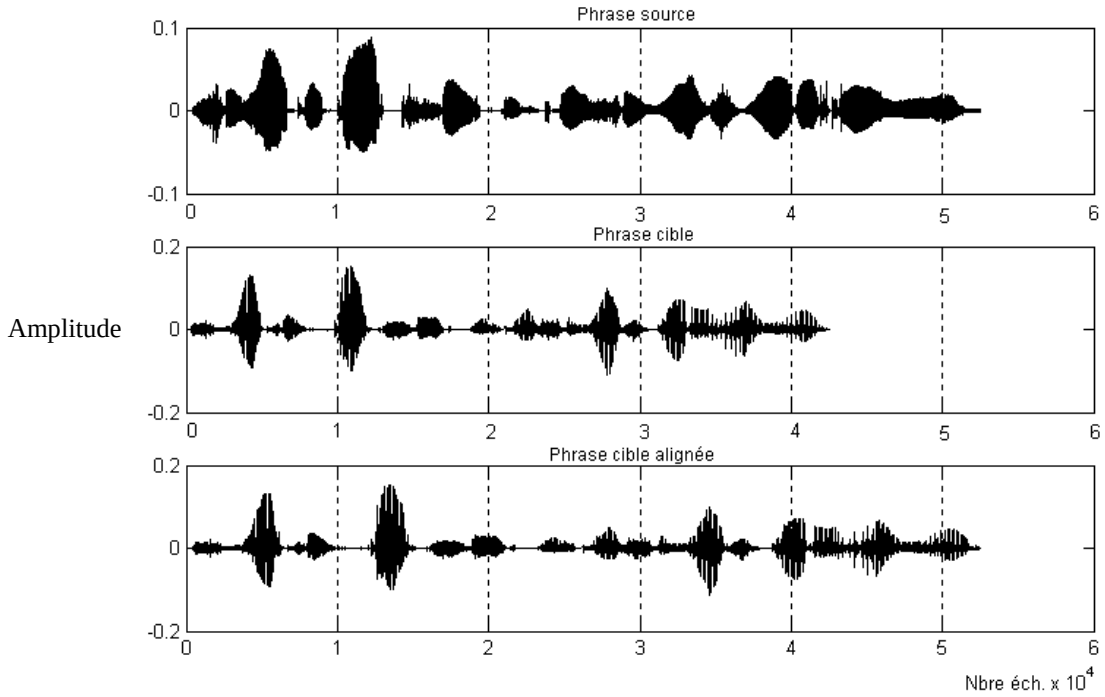

Figure 5.7 b « Alignement » de phrases par la méthode PSOLA
(silences avant et après la phrase supprimés)

5.2.2.3 « Mapping » par Régression linéaire

L'étape de mise en correspondance ou « mapping » représente une phase essentielle dans le processus d'apprentissage. Elle est utilisée dans une première étape pour déterminer les paramètres de transformation de la fonction de transfert du signal de parole source vers celle du signal de parole cible à travers les coefficients de prédiction. Celle –ci est généralement effectuée à travers une régression linéaire au sens des moindres carrés.

Nous avons cherché à améliorer les résultats en utilisant plutôt la régression linéaire « robuste » illustrée par la figure 5.8 basée sur une pondération permettant de suivre la ligne de régression de façon plus fidèle que par une simple régression.

La régression robuste a été testée sur les coefficients LSP et nous avons remarqué que certains coefficients étaient approchés de façon identique par les deux régressions alors que pour d'autres il y avait une différence due à la présence de valeurs particulières modifiant la pente globale de régression.
Nous avons finalement opté dans notre travail pour l'utilisation de la régression robuste.

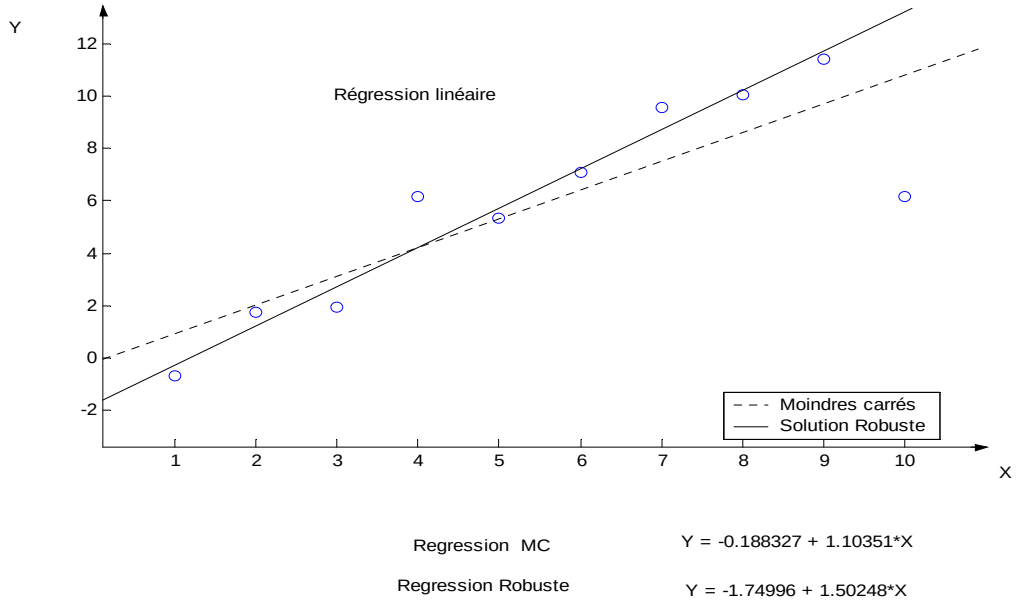

Figure 5.8 Illustration régression robuste

5.2.2.4 Mesure de performance.

Les coefficients LSP définis plus haut obtenus après alignement sont designés par a_srce_dtw et a_cble_dtw ce qui nous permet de définir l'erreur inter_locuteurs par :

$$E_{LSF}(srce,cble)=\frac{1}{M}\sum_{m=1}^{M}\sqrt{\frac{1}{P}\sum_{i=1}^{P}(a\_srce\_dtw^{m,i}-a\_cble\_dtw^{m,i})^2}$$

où M est le nombre de trames, P l'ordre LPC et a_…_dtw ^{m,I} est le I $_{ème}$ composant du vecteur LSF de la trame m.

Les coefficients LSP prédits après « mapping » seront notés a_new_dtw ce qui nous permet de définir l'erreur de transformation par :

$$E_{LSF}(cble,new)=\frac{1}{M}\sum_{m=1}^{M}\sqrt{\frac{1}{P}\sum_{i=1}^{P}(a\_cble\_dtw^{m,i}-a\_new\_dtw^{m,i})^2}$$

l'indice de performance IP sera alors donné par :

IP = 1- E(cble,new)/E(srce,cble).

5.2.2.5 Résultats et interprétation.

La base de données de parole TIMIT (Annexe A) contient deux phrases émises par tous les locuteurs de la base. Ce sont les phrases :

- SA1 : « She had your dark suit in greasy wash water all year ».
- SA2 : « Don't ask me to carry an oily rag like that ».

Elles sont échantillonnées à 16 Khz

Pour évaluer notre transformation, nous avons choisi dans cette base deux locuteurs males et deux locutrices femelles du répertoire DR1 correspondant a la région de New England des états unis. (voir tableau 5.9)

Locuteur	Référence	Durée Sa1 (s)	Durée Sa2 (s)
M1	Medr0	3.98	2.29
M2	Mklw0	3.16	2.69
F1	Felc0	3.77	2.96
F2	Fdrw0	3.25	2.52

Tableau 5.1 Locuteurs et durées des phrases émises.

Chacun des locuteurs est tantôt source tantôt cible selon le type de conversion expérimentée.

Le tableau suivant regroupe les résultats obtenus pour les différents types de conversion sur la base de la paire de phrase d'apprentissage Sa1 et Sa2.

Loc.Source	Loc Cible	Erreur Interlocuteur	Erreur de transformation	Indice de performance
F1	F2	0.018	0.012	0.298
M2	M1	0.016	0.008	0.485
F1	M1	0.018	0.011	0.401
F1	M2	0.021	0.010	0.511
M1	F1	0.018	0.009	0.495
M1	F2	0.019	0.011	0.420
F2	F1	0.018	0.012	0.324

Tableau 5.2 Erreurs inter-locuteurs et indice de performance
Ordre Lpc :12

Interprétation :

Il ressort à la lecture de ces résultats que la performance de la transformation diffère de façon significative selon la conversion envisagée, de même que la conversion pour les mêmes locuteurs, selon qu'ils sont source ou cible, aboutit à une performance différente. Ceci montre toute la complexité d'une évaluation objective en conversion de voix bien qu'elle soit généralisée dans toute étude de transformation de signal de parole. Les tests subjectifs jouissent d'ailleurs d'une plus grande importance pour les systèmes de conversion de voix lorsque la qualité des signaux synthétisés permet une bonne estimation des procédures de conversion.

La procédure d'alignement adoptée pourrait être à l'origine de ces écarts.

Ordre de l'analyse LPC

L'ordre de l'analyse LPC représentant une variable de notre système de conversion a fait l'objet de l'expérimentation dont les résultats sont consignés dans le tableau suivant pour une transformation femelle-femelle :

Ordre LPC	Erreur Interlocuteur	Erreur de transformation	Indice de performance
12	0.0176	0.0124	0.2980
14	0.0156	0.0109	0.2987
16	0.0138	0.0097	0.2994
18	0.0126	0.0088	0.3021
20	0.0115	0.0079	0.3167
22	0.0107	0.0073	0.3174

Tableau 5.3 Erreurs de transformation et indice de performance
Conversion F1-F2

La tendance des valeurs de performance montre qu'une valeur de l'ordre LPC égale à 20 représente une bonne performance, ce résultat est identique à presque toutes les conversions expérimentées et rejoint les résultats trouvés dans la littérature traitant du sujet [Kai01], [Gil03]. Ces résultats sont aussi en accord avec la base objective de l'évolution de l'énergie résiduelle de prédiction avec l'ordre qui indique une stabilisation au de la d'un ordre supérieur à 14 [BoK87].

Par ailleurs on estime que la transmittance du modèle doit comporter une paire de pôles par Khz de bande passante, l'excitation glottique et la radiation des lèvres exigeant ensemble 3 à 4 pôles. La différence entre les indices de performance correspondant aux valeurs extrêmes expérimentées n'étant pas très grande (à comparer aux différences observées sur le tableau 5.10) nous avons opté dans notre travail pour un ordre 12 classiquement utilisé et représentant un bon compromis entre coût de calcul et qualité de synthèse en accord avec la bande de fréquence du signal vocal.

Fréquence échantillonnage

Les résultats précédents ont été obtenus sur des signaux de parole échantillonnés à 16 Khz mais qui ont été rééchantillonés à 8 Khz .

Nous procédons maintenant à l'évaluation des indices de performance pour une fréquence d'échantillonnage de 16Khz

Loc.Source	Loc Cible	IP 8Khz	IP 16Khz
F1	F2	0.298	0.481
M2	M1	0.485	0.602
F1	M1	0.401	0.380
F1	M2	0.511	0.410
M1	F1	0.495	0.385
M1	F2	0.420	0.468
F2	F1	0.324	0.324

Tableau 5.4 Indices de performance
Ordre Lpc :12 Fe : 8 Khz et 16Khz

Interprétation

Globalement les indices de performance à 16 Khz sont supérieurs aux indices à 8 Khz, il reste néanmoins des exceptions (exple : conversion F1-M1 et F1-M2) pour lesquelles la seule explication qu'on a pu donner est liée aux caractéristiques phonétiques des phonèmes contenus dans les phrases utilisées pour nos expérimentations.
En effet la modélisation AR utilisée, connue comme étant une approximation du conduit vocal, n'est pas apte à une représentation efficace des sons fricatifs ou bien les nasales (le modèle ARMA étant plus approprié mais plus délicat à estimer); le fait d'augmenter la fréquence d'échantillonnage permet donc de compenser la partie MA (zéros de la fonction de transfert). Une étude plus approfondie sur les caractéristiques phonétiques des phrases utilisées pourrait nous renseigner sur la corrélation fréquence d'échantillonnage et indice de performance.

5.2.3 Transformation du contour de pitch.

5.2.3.1 Introduction

Le second paramètre auquel nous nous sommes intéressés dans notre élaboration du système de conversion est le pitch et les méthodes de sa transformation.
Dans les systèmes de conversion de voix rencontrés dans la littérature la transformation du pitch va de la simple mise à l'échelle respectant le pitch moyen du locuteur cible à une véritable prédiction.

Dans notre travail nous avons expérimenté différentes techniques nous permettant d'arriver à une qualité acceptable de conversion de voix.

5.2.3.2 Extraction des paramètres.

Plusieurs méthodes d'estimation de la période du fondamental et décision de voisement existent dans la littérature comme décrit dans le chapitre 4.

Nous avons utilisé dans notre travail la méthode du cepstre sur des trames de 128 échantillons.

- **Correction des valeurs erronées du pitch**

Les valeurs de pitch obtenues sont ensuite soumises à un post- traitement permettant de conserver un certain naturel au contour de pitch en éliminant certaines valeurs de pitch indésirables qui sont généralement observés (le pitch détecté est soit le double de la valeur correcte, un signal de période T est aussi de période 2T, soit la moitié et il s'agit dans ce cas de la 1ère harmonique.

Ceci est effectué par les procédures suivantes :

- Si une trame non voisée apparaît entre deux trames voisées la valeur de pitch de la trame est interpolée.
- Si une trame voisée apparaît entre deux trames non-voisées, la trame est considérée comme non-voisée et la valeur de pitch correspondante est forcée à zéro.
- Les valeurs de pitch sur les trames voisées sont finalement passées par un filtre médian apportant une amélioration supplémentaire au naturel du contour de pitch. (voir la figure 5.9 pour une illustration du résultat)

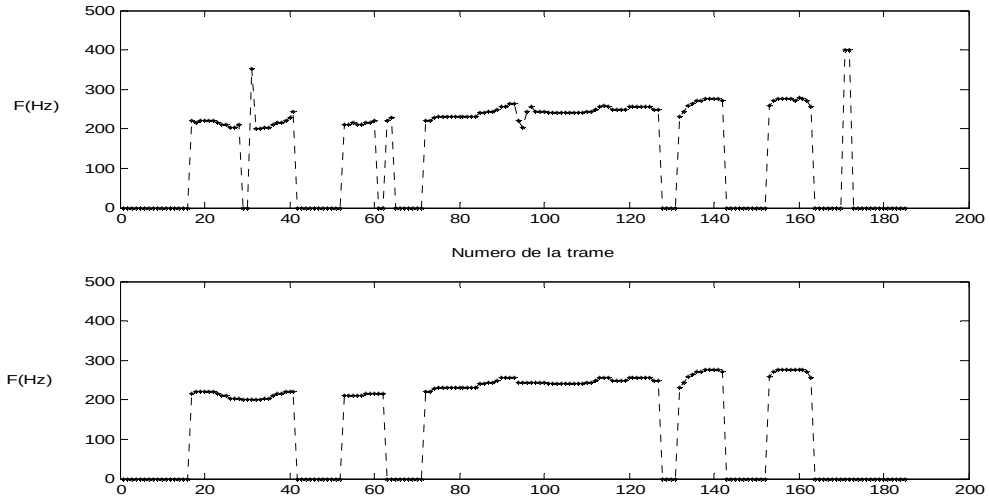

Figure 5.9 - Contour de pitch avant et après post traitement.

5.2.3.3 Conversion du contour de pitch

Plusieurs approches ayant pour but la conversion du contour de pitch ont été expérimentées :

- Transformation par simple mise à l'échelle par le rapport de la moyenne du pitch source à celle du pitch cible.
- Transformation par régression linéaire.
- Régression par la méthode moyenne-variance.

La première technique se résumant à une translation du contour de pitch source (à travers la phrase test) ne correspond pas à notre objectif de conversion. (contour noté Moy sur la figure 5.10 représentant une illustration des différentes méthodes).

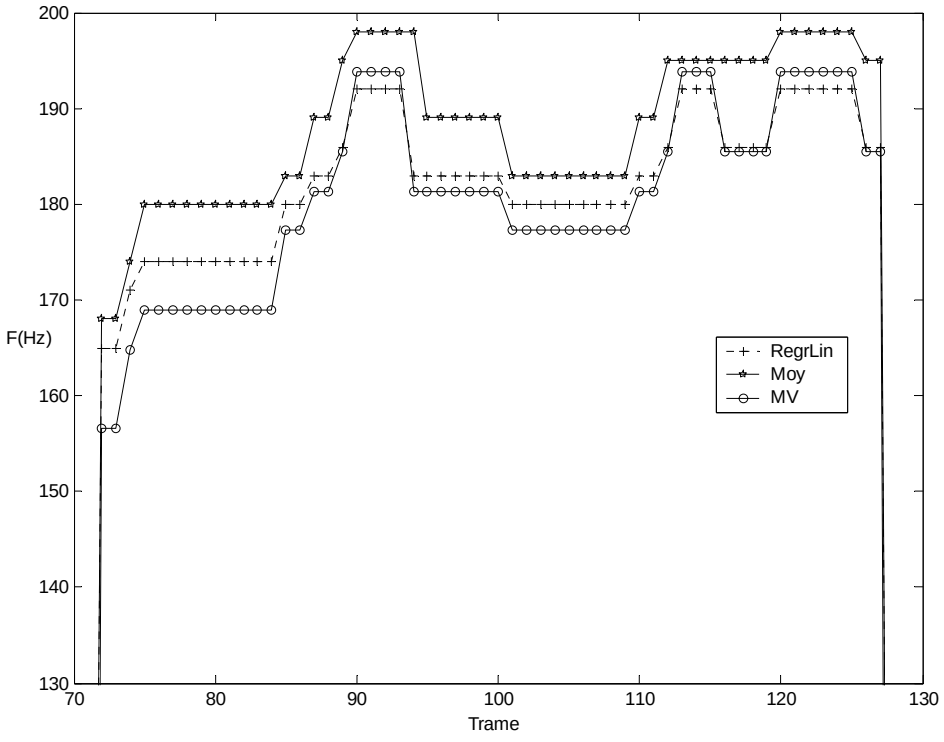

Figure 5.10 - Contour de pitch modifié

Quant à la deuxième méthode (contour noté RegLin), un biais correspondant au deuxième paramètre de la régression, selon $P_{modifié} = A . P_{test} + B$ est rajouté par rapport à la première formulation.

Malgré cela cette transformation n'a pas apporté d'amélioration perceptive notable. En effet ces deux approches sont basées sur une transformation linéaire destinée à respecter le pitch moyen d'un locuteur cible mais ne permettent pas de refléter des différences notables entre les formes de contour source et cible. Ceci peut s'expliquer par le fait que la phase d'apprentissage de laquelle sont issus les paramètres de la régression, (A et B) dans l'expression précédente, est basée sur une paire de phrases seulement.

La troisième approche décrite dans le chapitre 3 par contre nécessite un « apprentissage » à travers les paramètres statistiques calculés sur l'ensemble des phrases disponibles sur la base de données et pour chacun des locuteurs.
On voit bien sur cette figure que le contour noté MV (pour moyenne variance) ne correspond plus à un simple décalage sur l'axe des valeurs de pitch mais possède une dynamique qui non seulement respecte la valeur moyenne cible mais semble aussi suivre la prosodie cible à travers la variance obtenue sur l'ensemble des phrases d'apprentissage.
Les tableaux suivants résument les valeurs utilisées pour mener cette conversion concernant un locuteur male et un locuteur femelle dont les phrases prononcées ont fait l'objet de tests de conversion.

Locuteur1 : Felc0 de sexe féminin				
Réf	**Phrase**	**Durée**	**Moy.Pitch**	**Ecart Type**
Sa1	She had your dark suit in greasy wash water all year	3,80	213,04	29,90
Sa2	Don't ask me to carry an oily rag like that	2,96	208,94	38,44
Si1386	In wage negotiations, the industry bargains as a unit with a single union	5,52	205,92	22,54
Si2016	Heave on those ropes; the boat's come unstuck	3,40	214,63	35,08
Si756	Materials: ceramic modeling clay: red, white or buff	4,19	216,03	32,90
Sx 36	Only the most accomplished artists obtain popularity	3,52	211,68	24,92
Sx126	Artificial intelligence is for real	2,91	212,51	31,54
Sx 216	The small boy put the worm on the hook	2,20	208,24	21,64
Sx 306	A chosen few will become Generals	2,50	213,87	22,54
Sx396	The fish began to leap frantically on the surface of the small lake	3,82	197,42	35,39
	Moyennes sur l'ensemble des phrases		**210,23**	**29,49**

Tableau 5.5 Données statistiques locuteurs Felc0.

Locuteur 2 Medr0 sexe masculin				
Réf	**Phrase**	**Durée**	**Moy.Pitch**	**Ecart Type**
Sa1	She had your dark suit in greasy wash water all year	2,98	130,69	15,62
Sa2	Don't ask me to carry an oily rag like that	2,30	121,41	16,48
Si744	Get copper or earthenware mugs that keep beer chilled or soup hot	3,78	144,52	37,36
Si1374	This leaves the ordering of entries variable	2,27	121,23	19,80
Si2004	She had jumped away from his shy touch like a cat confronted by a sidewinder	4,12	128,40	23,98
Sx24	Although always alone, we survive	2,29	125,88	13,74
Sx114	The cartoon features a muskrat and a tadpole	2,53	145,03	32,44
Sx204	Even I occasionally get the Monday blues!	2,26	118,89	8,76
Sx294	Should giraffes be kept in small zoos?	2,00	132,39	11,39
Sx384	Please shorten this skirt for Joyce	1,95	121,82	25,49
	Moyennes sur l'ensemble des phrases		**129,03**	**20,43**

Tableau 5.6 Données statistiques locuteurs Medr0

La méthode « moyenne-variance », présentée au chapitre 3, adaptée à notre schéma de conversion à été appliquée comme suit :

$P_{predit} = A \cdot P_{test} + B$

Où $A = (var_{cible} / var_{source})^{1/2}$

$B = \mu_{cible} - (A \cdot \mu_{source})$

avec P_{test} : pitch de la trame courante de la phrase test

et var_{srcet} , var_{cible} les variances du pitch sur toute la base.

alors que μ_{source} et μ_{cible} sont les moyennes du pitch source et cible.

On aura alors :

$$P_{predit} = \left(\frac{var_{cible}}{var_{source}} \right)^{1/2} . P_{test} + \mu_{cible} - \left(\frac{var_{cible}}{var_{source}} \right)^{1/2} . \mu_{source}$$

5.2.3.4 Résultats et interprétation

Les valeurs du pitch moyennées sur l'ensemble des phrases de la base ainsi que l'écart type pour chaque locuteur sont donnés par les tableaux du type 5.13 et 5.14
Ceci nous permet d'évaluer les paramètres de la régression moyenne- variance pour chaque conversion ainsi que les valeurs de DPN mesurant la performance de la conversion.

Le tableau suivant illustre les résultats obtenus pour une conversion inter-genre.

Loc.Source	Loc Cible	Moyennes		Ecart type		Regression MV	
		Source	Cible	Srce	Cble	A	B
M1	F1	**129,03**	**210,23**	**20,43**	**29,49**	1,44	23,98
F1	M1	**210,23**	**129,03**	**29,49**	**20,43**	0,69	-16,03

Tableau 5.7 Résultats d'une conversion inter-genre .par MV (Phrase Sa1 pour apprentissage)

Ces résultats sont à comparer à ceux obtenus par les deux autres méthodes

Loc.Source	Loc Cible	Moyennes		Rap.Moy	Régression	
		Source	Cible		A	B
M1	F1	**130,69**	**213,04**	1,63	1,36	8,80
F1	M1	**213,04**	**130,69**	0,61	0,51	1,31

Tableau 5.8 Résultats d'une conversion inter-genre par régression (Phrase Sa1 pour apprentissage)

Conclusion

Ces modifications de pitch restent globales, en ce sens qu'elles ne s'appliquent qu'a des caractéristiques du pitch définies sur l'ensemble de la base de données analysée. Par conséquent, elles ne permettent pas de refléter des différences de style prosodique entre les locuteurs source et cible.

Une approche dans ce sens à été expérimentée pour pallier aux insuffisances de la régression linéaire quant à la prosodie, elle est exposée dans le paragraphe suivant .

Dans cette approche on a procédé comme suit :

- La courbe représentant le contour de pitch du locuteur cible est interpolée comme l'indique la figure 5.11

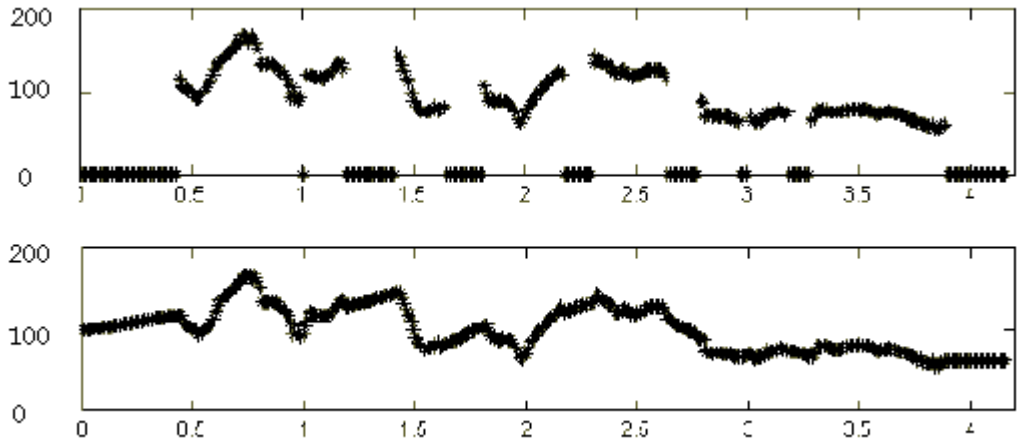

Figure 5.11 – Interpolation du contour de pitch sur les régions non voisées.

- On effectue ensuite un fenêtrage par la courbe de voisement de la phrase test que l'on veut convertir. On obtient ainsi le nouveau contour de pitch qui sera utilisé pour synthétiser le signal d'excitation.

Cette façon de procéder pour la conversion de pitch a pour avantage de donner à la phrase synthétisée une prosodie plus proche de celle du locuteur cible par rapport à celle obtenue par régression linéaire sous ses différentes formes. Néanmoins, cette méthode si elle est satisfaisante pour « copier » la prosodie de la phrase cible ne permet pas une véritable prédiction de style de prosodie.

L'estimation de caractéristiques plus locales relatives au pitch est beaucoup plus délicate. Une façon d'aborder le problème est de le reformuler en terme d'apprentissage prosodique. L'objectif est alors de relier des informations de type linguistique a des contours de pitch. Ce domaine de recherche a été et demeure toujours à l'origine de nombreux travaux. En synthèse de la parole notamment, l'apprentissage prosodique a pour but de prédire, en fonction du texte à synthétiser, les consignes qui permettront d'associer au message vocale une certaine prosodie. Certes, ces techniques automatiques de prédiction permettent de faire ressortir des caractéristiques prosodiques importantes, mais elles ne parviennent tout de même pas à restituer fidèlement la prosodie d'un locuteur.

5.2.4 Conversion par la méthode LP-PSOLA.

Le schéma de la conversion du pitch par la méthode LP-PSOLA est diffèrent du schéma global de conversion du pitch utilisé précédemment.

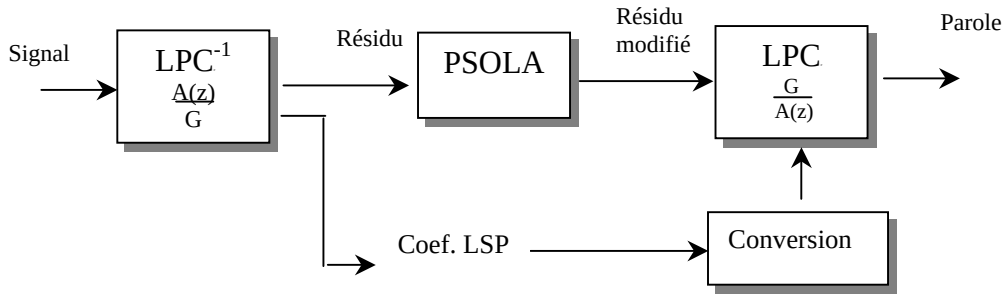

Figure 5.12 - Schéma synoptique de la transformation LP-PSOLA

Dans le cadre de la conversion de voix nous avons dû adapter le schéma de transformation LP-PSOLA à notre schéma global de conversion de voix (figure 5.4) comme suit :

- La conversion de l'enveloppe spectrale est effectuée par régression linéaire comme précédemment.
- Le signal résiduel issu de l'analyse LPC est converti par la méthode TD PSOLA .
- Le signal résiduel transformé et ensuite associé aux coefficients de prédictions convertis pour la synthèse du signal de parole converti.

Le filtrage inverse (analyse LPC) et la resynthèse sont des opérations linéaires alors que la transformation par TD-PSOLA est un processus non-linéaire la procédure est différente de l'équivalent dans le domaine temporel.

La procédure suivie pour l'implantation d'un tel schéma de transformation à été conditionnée par la modélisation par prédiction linéaire qui nous impose de considérer le pitch comme étant un paramètre segmental, la transformation s'effectuant trame par trame.

La méthode de transformation qui a été expérimentée comporte les étapes suivantes :

- Pour chaque trame du signal une analyse est effectuée (détermination du pitch et caractère de voisement et marquage) sur la base de cette analyse on procède à la conversion de l'enveloppe spectrale à travers une régression sur les coefficients de prédiction.
- La première et la dernière trame du signal de parole sont considérées comme non voisées et sont recopiées telle quelles.
- Pour chaque trame voisée on détermine la position des impulsions du pitch. Le marquage étant effectué par l' algorithme de Goncharoff exposé dans le chapitre 4.
- On applique la transformation TD-PSOLA sur le résidu de la phrase test en supprimant ou dupliquant les segments fenêtrés obtenus dans l'étape précédente pour obtenir le nouveau signal résiduel de la trame courante. Les figures 5.13 et 5.14 illustrent ces transformations sur le résidu.

- Pour une trame non voisée on génère un bruit blanc conformément au schéma conventionnel du modèle de FANT (voir figure 5.5)
- La resynthèse est alors effectuée à base du résidu transformé.

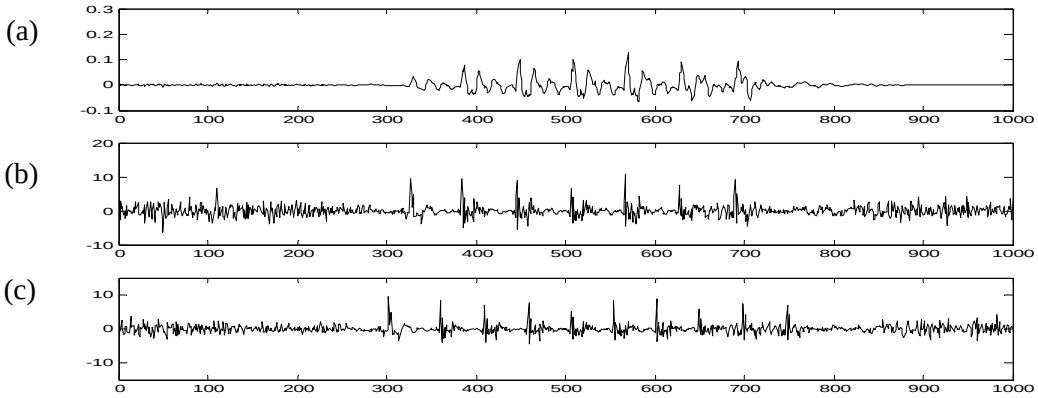

Figure 5.13 - Modification du signal résiduel par la méthode PSOLA
(a) signal (b) résidu (c) résidu avec un pitch augmenté.

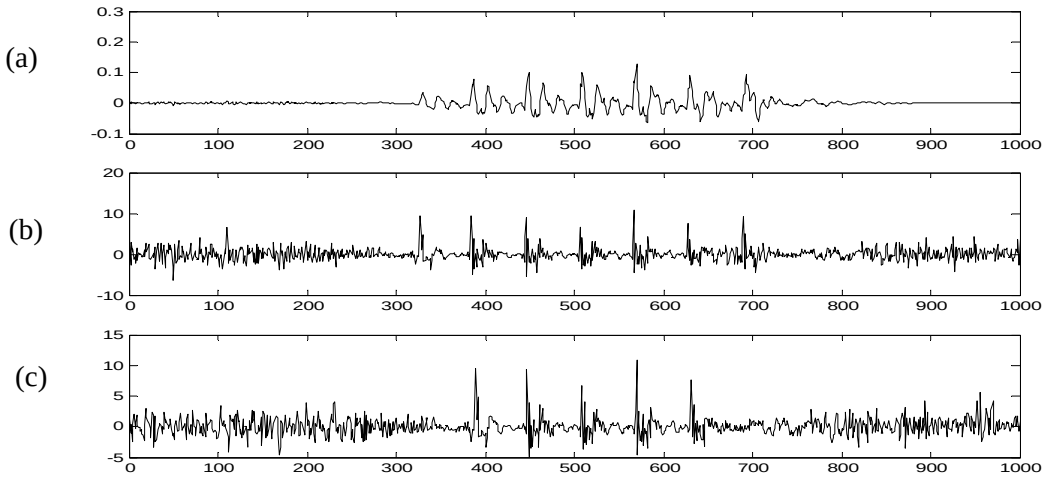

Figure 5.14 - Modification du signal résiduel par la méthode PSOLA.
(a) signal (b) résidu (c) résidu avec un pitch réduit

A travers les tests d'écoute effectués sur les signaux synthétisés cette manière d'appliquer la méthode PSOLA à la conversion de voix n'a pas apporté d'amélioration perceptible par rapport à la méthode moyenne-variance et cela en faisant abstraction des problèmes de continuité d'une trame à l'autre qui on induit des artefacts désagréables.

Il fallait donc à ce niveau de notre mis au point de la méthode déceler l'origine de ces artefacts. L'expérience suivante nous a permis de les localiser et modifier la procédure.

Ceci est illustré par les spectrogrammes de la figure 5.15 relatifs respectivement au signal synthétise avec l'excitation issue de l'analyse LPC, puis avec un bruit blanc comme excitation et enfin celui de l'excitation originale seule.

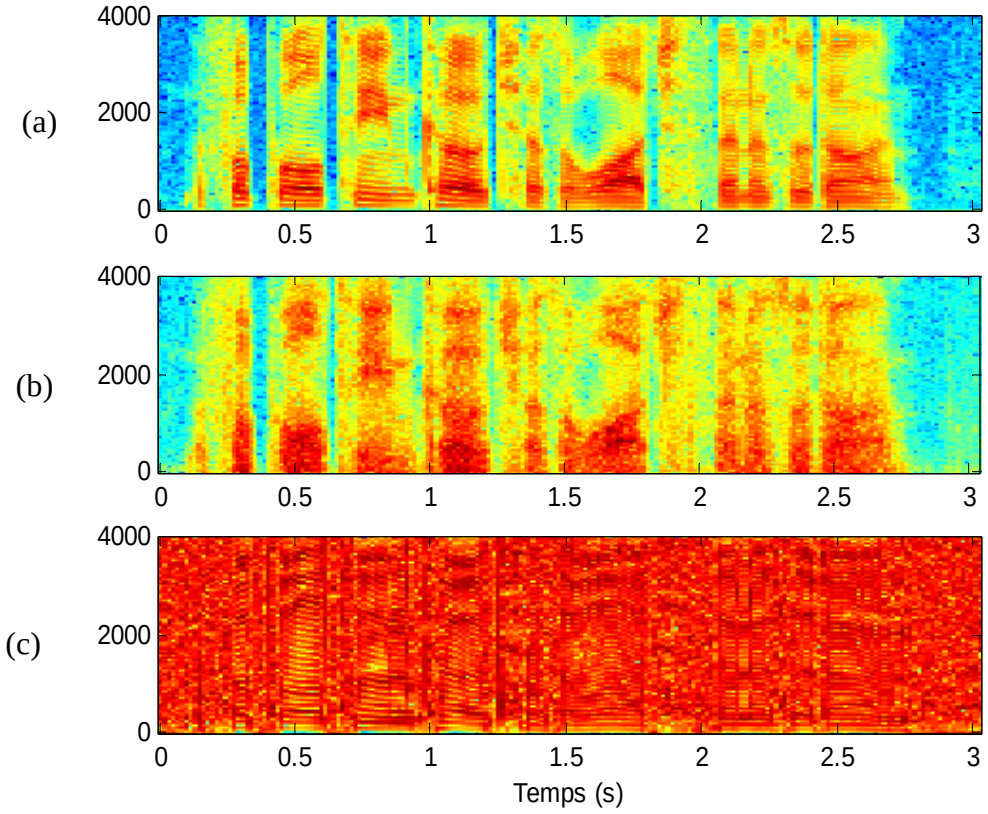

figure 5.15 – Spectrogrammes de signaux synthétisés
(a)avec excitation (b) avec un bruit blanc (c) signal d'excitation

On voit bien que les deux spectrogrammes (a) et (b), ont la même forme mais la structure de pitch est absente du second. Le test d'écoute correspondant au second spectrogramme à base de bruit blanc comme excitation nous a donné l'impression d'une « voix enrouée » car dénuée de pitch.
Dans le spectrogramme du bas représentant le résidu seul on voit bien que les formants sont encore visibles et son écoute est encore intelligible bien que la phrase est noyée dans du bruit.

Les conclusions qu'on à pu tirer de cette simple expérience sont multiples :

- Lorsque des artefacts de ce type se produisent dans les phrases converties leur origine est alors bien cernée et cela permet de mieux envisager leur élimination ou du moins la diminution de leurs effets.

- D'autre part, le résidu est porteur d'informations supplémentaires concernant l'identité du locuteur et ce en dehors de l'information du pitch.

- L'utilisation du résidu source n'est donc pas recommandé et explique les effets de réminiscence de la voix du locuteur source dans certains cas de conversion.

Cette expérience nous a conduit à la modification suivante sur la procédure précédente :

Pour les trames non voisées au lieu d'utiliser un bruit blanc comme excitation, on introduit le résidu cible qu'on aura au préalable modifié par la méthode PSOLA celle – ci permettant les modifications temporelle et fréquentielle.

Les test d'écoute effectués à la suite de cette modification nous ont permis d'observer une nette amélioration du naturel des signaux synthétisés.

Interprétation

Il est connu que les procédures de transformation de l'enveloppe spectrale (par exemple dans les systèmes de codage) utilisant les LSF de manière générale ont tendance à atténuer les détails spectraux qui contribuent au naturel des signaux de parole cet inconvénient peut donc être contrecarré par l'utilisation du résidu pour améliorer la qualité des signaux synthétisés.

Evaluation subjective des résultats

Pour évaluer notre conversion sur le plan perceptif, nous avons procédé à des tests d'écoute. Le test d'écoute le plus utilisé pour évaluer un système de conversion de voix est le test ABX : c'est un test dans lequel des auditeurs jugent si la phrase prononcée par un locuteur « X » est plus proche du locuteur « A » ou de « B ».
Quant à la qualité globale de la conversion celle-ci est plutôt évaluée par un test MOS (Mean Opinion Score) sur une échelle à cinq niveau qui va de la valeur un, pour la plus mauvaise qualité, jusqu'à cinq pour une qualité excellente.

Plusieurs tests de conversion ont été réalisés pour des locuteurs de même sexe de même que pour des locuteurs de sexe différent sur plusieurs phrases issues de la base TIMIT.

Les résultats obtenus ont fait l'objet d'une évaluation subjective à travers des tests d'écoute parmi une population avertie (collègues du laboratoire traitement du signal et de la parole) et non avertie (entourage).

Les test d'écoute on porté d'une part sur l'intelligibilité des phrases converties, d'autre part sur la qualité de la conversion, en faisant abstraction de la qualité des phrases synthétisées celle-ci ayant fait l'objet d'une évaluation séparée.

En général, les phrases converties sont parfaitement intelligibles ce qui dénote qu'il n'y a pas eu de distorsions significatives induisant une dégradation de l'intelligibilité des phrases utilisées lors des différentes transformations.
La perception du genre obtenue à été dans l'ensemble convenable pour la majorité des conversions effectuées.
La qualité des phrases converties était par contre moyenne et dépendait parfois des phrases utilisées alors que pour certains tests il restait dans la phrase convertie des réminiscences de la phrase source et dont l'origine à été expliquée plus haut.

Les valeurs moyennes des résultats obtenus sont reportées sur le tableau suivant :

Conversion / Test	Male vers male	Male vers femelle	Femelle vers male	Femelle vers femelle
ABX1	75%	80%	80%	60%
MOS1	5	5	4.5	5
MOS2	3	2.5	2.4	3

Tableau 5.9 Résultats des tests subjectifs

ABX1 : Conversion

Mos1 : Intelligibilité

Mos2 : Qualité des phrases

Interprétation

Notons que ces résultats ont été obtenus sur la base d'un nombre limité de phrases et de locuteurs (deux males et deux femelles de la même région voir Tableau 5.9), de ce fait les résultats obtenus sont loin des recommandations connues à ce sujet et suivant un protocole rigoureux; néanmoins et dans le cadre de la mise en place de la méthodologie de conversion de voix ces résultats peuvent être pris en considération et reflètent parfaitement les capacités du système élaboré à convertir de façon très convenable une phrase « test » à travers un apprentissage sur la base d'une paire de phrase (source et cible) de durée limitée.

Dans le cas de la transformation du contour de pitch par la méthode moyenne variance on peut estimer la durée de l'apprentissage à une trentaine de minute, ce qui représente une durée convenable d'après la littérature.

Conclusion et perspectives

A l'issu de ce travail nous estimons avoir atteint notre objectif qui consistait à mettre en place un système d'analyse - transformation - synthèse de parole.

Le système mis en place nous a permis d'étudier les transformations qui permettent de donner l'identité de la voix d'un locuteur « cible » à celle d'un locuteur « source ».

La méthodologie suivie devait utiliser l'extraction, l'apprentissage et la modification de paramètres du signal vocal (fréquence fondamentale, formants, excitation, etc.) liés au locuteur. Dans cette étude, un système simplifié a été développé en utilisant la modélisation par prédiction linéaire et la technique PSOLA et des résultats préliminaires obtenus.

Ce travail nous a permis de cerner les problèmes relatifs à l'alignement temporel, la transformation de l'enveloppe spectrale et la prédiction du contour de pitch qui sont considérés comme les problématiques principales d'un système de conversion de voix. Conversion qui ne peut être efficace sans une analyse et une synthèse de qualité.

De ce fait les améliorations pouvant être apportées à ce système de base à travers une amélioration de l'alignement temporel comme premier pas du futur système ainsi qu'une prise en compte dans les transformations envisagées d'un modèle d'apprentissage des différentes composantes prosodiques, devrait permettre des améliorations perceptives très importantes. Par ailleurs l'introduction de méthodes plus robustes au niveau de l'analyse (particulièrement la détection de pitch) et la synthèse par l'utilisation de la technique PSOLA qui est caractérisée par la très bonne qualité des signaux de parole qu'elle génère, représentent autant de sujets d'intérêt pour des axes de recherche futurs.

La base de données TIMIT.

La base de données utilisée dénommée :
« DARPA TIMIT Acoustic-Phonetic Speech Database »
est une production de l'Institut National des Standards et Technologie (NIST).
Elle est disponible sous forme de CD et comprend un ensemble de 420 locuteurs et 4200 phrases en anglais selon différentes régions des Etats Unis d'Amérique.

La base est organisée comme suit :

- Chaque répertoire correspond à l'identificateur du dialecte régional selon le tableau suivant :

Identificateur	Region
dr1	New England
dr2	Northern
dr3	North Midland
dr4	South Midland
dr5	Southern
dr6	New York City
dr7	Western
dr8	Army Brat

- L'identificateur du locteur est defini comme suit : Ex: mrws0

Caractère	signification
M	Sexe du locuteur
rws	Initiales du locuteur
0	N° du locuteur *

* pour les locuteurs ayant les mêmes initiales.

- Chaque sous-répertoire de locuteur comprend 10 phrases Ex: sa1

Caractère	signification
sa	Type de phrase
1	N° de la phrase

Le type de phrase est défini comme suit :

sa – phrase de calibration (SRI) (2 par locuteur)
si – phrase à contexte aléatoire variable (TI) (3 par locuteur)
sx - phrase phonétiquement équilibrée (MIT) (5 par locuteur)

Chaque répertoire de phrase contient trois fichiers connexes relatifs à la phrase.
Les types de fichier sont :

Le fichier « *. adc » contient une entête type CMU suivie par les données paroles numerisées.
Le fichier « *. phn » contient la transcription phonétique de la phrase.
Le fichier « *. txt » contient la transcription orthographique.

Bibliographie

[Abe88] M. Abe, S. Nakamura, K. Shikano, H. Kuwabara," Voice Conversion through Vector Quantization," Proc. ICASSP 1988, New York,USA, pp. 655-658.

[Abe91] M. Abe. A segment-based approach to voice conversion, Proc. of ICASSP, Toronto,Canada, May 1991, pp. 765-768.

[Ars99] L.M. Arslan “Speaker Transformation Algorithm using Segmental Codebooks (STASC)’’, Speech Communication Journal, vol. 28, pp. 211-226, June 1999.

[BaS96] G. Baudoin et Y. Stylianou, « On the transformation of the speech spectrum for voice conversion », In *Proc. of ICSLP* Philadelphia,PA, USA October 3-6,1996, pp.1405-1408.

[BoK87] R.Boite et M. Kunt, « Traitement automatique de la parole », Presses polytechniques romandes,1987.

[Che03] Y. Chen, M. Chu, J. Liu et R. Liu { « Voice Conversion with smoothed GMM and MAP adaptation », Proc. of Eurospeech Geneva, Switzerland: ISCA, Sept.2003:2413-2416.

[Enn05] T.Ennajary, « Conversion de voix pour la synthèse de la parole » , Thèse de Doctorat, Université de Rennes France. (2005).

[Fan66] G. Fant, « A note on vocal tract size factors and noninuform F-Pattern scalings », *Speech Transmission Laboratory Quaterly Pregress and Status Report 4/66, Royal Institute of Technology,Sweden* (1966).

[Fan70] G. Fant, « Acoustic theory of speech production », Mouton 1970 The Hague. (1970)

[Fan75] G. Fant, « Non-uniform vowel normalization », *Speech Transmission Laboratory Quaterly Pregress and Status Report 2-3/75, Royal Institute of Technology,Sweden* (1975).

[Gil03] B. Gillett. « Transforming voice quality and intonation », MSc. Thesis.University of Edimburgh. (2003)

[Hej90] D. J. Hejna Jr. Real-time time-scale modification of speech via the synchronized overlap-add algorithm. Master's thesis, Massachusetts Institute of Technology, USA Feb.1990. Department of Electrical Engineering and Computer Science.

[GoG98] Goncharoff, V. and P. Gries.”An Algorithm for Accurately Marking Pitch Pulses in Speech Signals”. In Proc. of the SIP’98, Las Vegas, USA. 1998.

[HeM91] Don Hejna and Bruce R. Musicus, "The SOLAFS Time-Scale Modification Algorithm", BBN Technical Report, July 1991.

[GrK90] J.Grass and P.Kabal, « Comparison of LPC coefficient quantizers », in Proc.Biennial Symp.Commun.(Kingston,ON),pp312-315, June 1990.

[Ina03] Z. Inanoglu, « Transforming Pitch in a Voice Conversion Framework », Master, St.Edmund's CollegeUniversity of Cambridge July , 2003

[IS88] K.Itoh et S.Saito,«Effects of acoustical features parameters on perceptual speaker identity », *Review of the Electrical Communications Laboratory* 36(1988),no.1, p.135-141.

[ITU96] International Telecommunication Union, "ITU-T Recommendation P.800: Methods for subjective determination of transmission quality," 1996.

[Kai01] A.Kain, *High resolution voice transformation*, Ph.D. Thesis, Oregon Health and Science University, Portland, OR, Oct. 2001.

[KaM98] Kain, A. and M. W. Macon..” Spectral Voice Transformations for Text-to-Speech Synthesis. In Proc. of the ICASSP’98, Seattle, USA. 1998.

[Kei03] Keizai Y. and Ikeda M « PICOLA and TDHS » Web page 2003 **http://keizai.yokkaichi-u.ac.jp/~ikeda/research/picola-jp.html**

[KM01] A. Kain, M. Macon. Design and evaluation of a voice conversion algorithm based on spectral enveloppe mapping and residual prediction, Proc. ICASSP 2001, Salt Lake City,USA, pp.813-816.

[LaC98] Y.Laprie and V.Colotte.”Automatic pitch marking for speech transformation via TD-PSOLA”. In IX European Signal Processing Conference, Rhodes, Greece, 1998.

[Mak75] J.Makhoul, « Linear prediction : A tutorial review », Proc.of the IEEE, vol 63, no 4 pp. 561-580, (1975)

[Mar00] Ph. Martin, "Peigne et brosse pour Fo : Mesure de la fréquence fondamentale par alignement de spectres séquentiels", *Actes des 23èmes JEP*, Aussois, France, juin 2000, 245-248.

[Mar72] J. D. Markel, A. H. Gray jr. Linear prediction of speech, Springer-Verlag. (1972)

[MaW79] [MW79] H. Matsumoto et H. Wakita, « Frequency warping for nonuniform talker normalization », *Proc. of ICASSP* (1979).

[MoC90] E. Moulines, F. Charpentier. Pitch-synchronous waveform processing techniques for text-to-speech synthesis using diphones, Speech Communication 1990.

[MG76] J. D. Markel and A. H. Gray, *Linear Prediction of Speech*, 1976.

[MHSN73] H. Matsumuto, S. Hiki, T. Sone et T. Nimura, «Multidimentional representation of personal quality of vowels and its acoustical correlates », *IEEE Trans.* AU-21(1973), p.428-436.

[Mou90] E. Moulines, « *Algorithmes de codage et de modification des paramètres prosodiques pour la synthèse de parole à partir de texte »*, Ph.D. thesis, Ecole Nationale Supérieur des Télécommunications (Paris), 1990.

[MV95] E. Moulines, w. Verhelst. “Time-Domain and Frequency-Domain Techniques for Prosodic Modification of Speech”,Speech coding and synthesis Chapitre 15 Elsevier Science 1995

[Nol75] P. Nordstrom et B. Lindblom, « A normalization procedure for vowel formant data », *International Congress of Phonetic Sciences* (1975).

[OrS01] Nicola Orio and Diemo Schwarz. « Alignment of Monophonic and Polypophonic Music to a Score ». In *Proceedings of the ICMC*, Havana, Cuba, 2001.

[PaK95] K.K. Paliwal and W.B. Kleijn “ Quantization of LPC Parameters”, in Speech Coding and Synthetis, Eds Amsterdam : Elsevier Science BV,1995

[Pal95] K. K. Paliwal, "Interpolation Properties of Linear Prediction Parametric Representations," Proceedings of Eurospeech, pp. 1029-32, September 1995.

[Pee01] G. Peeters, « Modèles et modification du signal sonore adaptés à ses caractéristiques locales », Thèse de doctorat, Université Paris VI. (2001).

[Pee02] G.Peeters, « Analyse et synthèse SINOLA », http://www.ircam.fr/anasyn/peeters/ . (2002)

[Por76] Portnoff, M.R. (1976) "Implementation of the Digital Phase Vocoder Using the Fast FourierTransform" *IEEE Transactions on Acoustics,Speech and Signal Processing*, ASSP-24, no 3, June 1976.

[Por81] Portnoff, M.R. (1981) “Time-scale modification of speech based on short-time fourier analysis”, *IEEE Transactions on Acoustics, Speech, and Signal Processing*, ASSP-29(3), 374-390, June 1981.

[Qua92] T. Quatieri and R. McAulay, “Shape invariant time-scale and pitch modification of speech ”, IEEE Trans. on Sig. Proc., vol. 40, no.3, pp. 497-510,1992.

[Ros74] M. J. Ross, H. L. Shaffer, A. Cohen, R. Freudberg and H.J.Manley, "Average Magnitude Difference Function Pitch Extractor," IEEE Trans. on Acoustics, Speech and Signal Processing, vol. 22, no. 5, pp. 353362, 1974.

[Sch98] D.Schwarz (1998). Spectral envelopes in sound analysis and synthesis, mémoire de Master, Université de Stuttgart.

[Sty95] Y. Stylianou, O. Cappé, E. Moulines (1995). Statistical methods for voice quality transformation, Proc. Eurospeech 1995,Madrid, Espagne, pp. 447-450.

[Sty96] Y. Stylianou, « Decomposition of speech signal into a deterministic and a stochastic part » Proc. Of ICSLP 1996.

[Son68] M. M. Sondhi, "New Methods of Pitch Extraction," IEEE Trans. on Audio and ElectroAcoustics, vol. 16, no. 2, pp. 262266, 1968.

[Sun02] Xuejing Sun « Pitch determination and voice quality analysis using subharmonic – to – harmonic ratio » in Proc. of ICASSP2002, Orlando, Florida, May 2002.

[Sün05] D. Sündermann, A. Bonafonte, H. Ney, and H. Höge, « A Study on Residual Prediction Techniques for Voice Conversion», in Proc. of the ICASSP'05, Philadelphia, USA, 2005.

[Tod04] T. Toda, H. Saruwatari et K. Shikano, « Voice conversion algorithm based on gaussian mixture model with dynamic frequency warping of STRAIGHT spectrum », *Proc. of ICASSP* (2004), no. I-97-I-100.

[Tur03] O.Türk, « New Methods for Voice Conversion », Ph.D. thesis, Bogaziçi University, Istanbul, Turkey. 2003.

[Val92] H. Valbret, E. Moulines, J.P. Tubach (1992). Voice transformation using PSOLA technique, ICASSP 1992, vol. 1, pp. 145-148.

[Ver00] W. Verhelst, (2000) “Overlap-add methods for time-scaling of speech”, *Speech Communication*, (30), 207-221.

[ViV05] M.Vondra, R.Vich, « Voice conversion based on spectral envelope transformation » in Lecture Notes in Computer Sciences, Vol.3445/2005

[VMT92] H.Valbret, E.Moulines et J.P.Tubach (1992), "Voice transformation using PSOLA technique", *Speech Communication, 11, n° 2/3, June 1992, pp. 175-187.*

[Yey03] H.Ye, and S.Young,” Perceptually Weighted Linear Transformations for Voice Conversion”, In Proc. of the Eurospeech’03, Geneva, Switzerland. 2003.

[YeY04] H. Ye and S. J. Young, « Voice Conversion for Unknown Speakers », in Proc: of the ICSLP'04, Jeju, South Korea, 2004.